1
INTRODUCTION

The social sciences today are witnessing a period of robust activity. Political scientists are deepening our understanding of the processes of cooperation and competition among nations, anthropologists are developing new tools for understanding the phenomena of culture—our own as well as that of others, historians are shedding new empirical and explanatory light on the past, and so on among the human sciences. Yet for the student of the social sciences—and perhaps for the practitioners themselves—one of the results of this growth is a sense of methodological cacophony. The issues and methods that are fundamental for one science are unknown to another; complex, many-sided disputes in one field are seen as arcane and pointless in another. Important questions arise. In what sense are the human sciences *scientific?* In what ways do they share empirical methods and explanatory paradigms? Is it possible to identify a coherent framework of assumptions about method, evidence, and explanation that underlies the practice of diverse social sciences?

In the strict sense the answer to this last question is no. There is virtually nothing in common between, for example, the thick descriptions of Balinese practices offered by Clifford Geertz and the causal analysis of English demographic change offered by Roger Schofield. However, in a looser sense there is room for some confidence that a degree of unification may exist—not around a single unified method of social inquiry but around *a* cluster of explanatory models and empirical methods employed in a wide range of social sciences today. Many social sciences offer causal explanations of social phenomena, for example; therefore it is important to clarify the main elements of the notion of social causation. Many social sciences premise their explanations on assumptions concerning the nature of human *agency*—both rational choice explanations and hermeneutic interpretations. Structural and functional explanations likewise play a role in a variety of social sciences, and issues concerning the microfoundations of macrophenomena crop up again and again in political science, economics, and sociology.

So there is a cluster of topics in the theory of explanation that together permit us to understand a wide range of social science research programs. The aim of this book is to examine the logical features of many of these topics. The level of detail is important. I have tried to avoid highly technical issues in order to make the discussion accessible to a wide audience. At

the same time I have striven for philosophical adequacy—to provide an account of these issues that is sufficiently nuanced to avoid traditional pitfalls and to shed light on current social science practice.

This book is organized around a large number of concrete examples of current social-scientific research. It contains discussion of social science explanations drawn from anthropology, geography, demography, political science, economics, and sociology; it also discusses social phenomena ranging from Asian peasant societies to patterns of residential segregation in industrialized societies. I have taken this approach because I believe that the philosophy of social science must work in close proximity to the actual problems of research and explanation in particular areas of social science, and it must formulate its questions in a way that permits different answers in different cases. Before we can make significant progress on the most general issues, it is necessary to develop a much more detailed conception of the actual models, explanations, debates, methods, etc., in contemporary social science. And we will need to develop a deeper recognition of the important degree of diversity found among these examples. I will therefore approach the general problems of the philosophy of social science from below, through examination of particular examples of social-scientific ex-planation. Close study of some of this diverse material will indicate that there is no single unified social science but rather a plurality of "sciences" making use of different explanatory paradigms and different conceptual systems and motivated by different research goals. Instead of a unity of science, a plurality of sciences will emerge.

In this view of the philosophy of social science, philosophers stand on the boundary between empirical research and purely philosophical analysis. Their aim is both to deepen our philosophical understanding of the social sciences through careful consideration of concrete research and theorizing and to provide the basis for progress in the area of science under consideration through careful analysis and development of the central theoretical ideas. Philosophers can learn about the logical structure and variety of social sciences only by considering specific examples in detail—thereby contributing to a more comprehensive theory of science that is genuinely applicable to social science. But at the same time philosophers can contribute to ongoing theoretical controversies in specific areas of research by clarifying the issues, by offering the results of other areas of philosophy (for example, rational choice and collective choice theory), by suggesting alternative ways of characterizing the theoretical issues, etc.'

PLAN OF THE BOOK

The chapters that follow are organized into three parts. Part I introduces three important ideas about the character of social explanation: that it requires identifying causes, that it flows from analysis of the decisionmaking of rational agents, and that it requires interpretation of culturally specific norms, values, and worldviews. These three ideas underlie much current social

explanation, and they provide the basis for many of the debates that arise in the philosophy of social science.

Part II turns to elaborations and combinations of these basic models of explanation. Functional and structural explanations are sometimes thought to be distinctive types of explanation, but Chapter 5 argues that each depends on causal explanation of social phenomena. Materialist explanation (Marxism and related theories) is sometimes believed to be an autonomous form of explanation as well, but Chapter 6 suggests that this model of analysis actually depends on both rational choice and causal explanations. Economic anthropology, discussed in Chapter 7, attempts to explain features of social behavior and organization of premodern societies on the basis of rational choice models; much of the debate in this field follows from the contrast between rational choice and interpretive explanations described in Part I. And statistical explanations, common in many areas of social science, are sometimes held to be more rigorous than other forms of explanation. Chapter 8 presents the central ideas of statistical explanation and concludes that it is in fact a form *of* causal explanation.

Part III turns to several general problems in the philosophy of social science that arise throughout the first two parts. Chapter 9 considers the topic of methodological individualism, Chapter 10 turns to the topic of cultural relativism, and Chapter 11 concludes with a discussion of the doctrine of naturalism as a methodology for social science.

Each chapter contains a number of examples of social science explanation. These examples are chosen to illustrate various aspects of such explanation and to give the reader a more concrete understanding of social science research. They have been separated from the text so that the reader can refer to them more conveniently.

SCIENTIFIC EXPLANATION

The main topic *of* this book is the nature *of* social explanation. But we need to pose one question before we can proceed to the details: What is a scientific explanation? Let us refer to the event or pattern to be explained as the explanandum; the circumstances that are believed to explain the event may be referred to as the *explanans.* (See Figure 1.1.) What is the relation between explanans and explanandum in a good explanation?

The topic of scientific explanation encompasses several different questions. What is the purpose of a scientific explanation? What is the logical form of an explanation? What are the pragmatic requirements of explanation? What are the criteria of adequacy of an explanation? And what role do general laws play in scientific explanations?

Explanans

Explanandum

Fig. 1.1 Logic of explanation

"Why" questions

Explanation usually involves an answer to a question. Why did the American Civil War occur? Why are two-party democracies more common than multiple-party democracies? Why is collectivized agriculture inefficient? How does the state within a capitalist democracy manage to contain class conflict? These questions may be divided into several different categories. Some may be paraphrased as "why-necessary" questions, and others may be described as "how-possible" questions. Consider the "why-necessary" question. Here the problem is to show that an event, regularity, or process is necessary or predictable in the circumstances—that is, to identify the initial conditions and causal processes that determined that the explanandum occurred. Here we are attempting to identify the sufficient conditions that produced the explanandum. This description is overly deterministic, however; in many cases the most that we can say is that the circumstances described in the explanans increased *the probability* of the occurrence of the explanandum.

Answers to "why" questions commonly take the form of causal explanations—explanations in which we identify the cause of a given outcome. But there are other possibilities as well for a "why" question may provoke explanation based on an agent's motivations. Why did the Watergate cover-up occur? Because the president wanted to conceal knowledge of the break-in from the public before the election. Here, then, the "why" question is answered through a hypothesis about the agent's motives. And a "why" question may invite functional explanation as well. Why do bats make squeaky noises? Because they use echolocation to identify and capture their prey. In this case the question is answered by reference to the function that the squeaky noise capacity plays in the bat's physiology.

The other central type of explanation-seeking questions is the "how-possible" question. Generally these concern the behavior of complex systems--complicated artifacts, neural networks, social organizations, economic institutions. We note a capacity of the system—say, the ability of a frog to perceive a fast-moving fly and catch it with a quick flick of the tongue—and then we attempt to produce an account of the internal workings of the system that give rise to this capacity. A market economy has the capacity to produce inputs in approximately the proportions needed for the next production period, and we may ask how this is possible—that is, what are the economic mechanisms that induce steel, rubber, and plastic manufacturers to produce just the right amounts to supply the needs of the automobile industry?

"How-possible" questions are related to the demand for functional explanations of parts of systems. In this case we need to provide a description of a functioning system in which various subsystems perform functions that contribute to the performance capacity that the larger system is known to have. These are in fact a species of causal explanation; we are attempting to discover the causal properties of the subsystems in order to say how these systems contribute to the capacity of the larger system.

The covering-law model

What is the logical structure of a scientific explanation? We may begin with a common view, the covering-law model, based on the idea that a given event or regularity can be subsumed under one or more general laws. The central idea is that we understand a phenomenon or regularity once we see how it derives from deeper regularities of nature. In other words, the event or regularity is not accidental but rather derives from some more basic general law regulating the phenomenon. The covering-law model thus takes its lead from this question: Why was the phenomenon to be explained *necessary* in the circumstances?

This insight has been extensively developed in the form of the *deductive-nomological* (D-N) model of explanation (Figure 1.2). According to this approach an explanation is a deductive argument. Its premises include one or more testable general laws and one or more testable statements of fact; its conclusion is a statement of the fact or regularity to be explained. Carl Hempel's classic article "The Function of General Laws in History" (1942) provides a standard statement of the D-N model of explanation. "The explanation of the occurrence of an event of some specific kind **E** at a certain place and time consists . . . in indicating the causes or determining factors of **E. . . .** Thus, the scientific explanation of the event in question consists of (1) a set of statements asserting the occurrence of certain events $\mathbf{C}_1$, $\mathbf{C}_2$, ... , $\mathbf{C}_n$ at certain times and places, (2) a set of universal hypotheses, such that (a) the statements of both groups are reasonably well confirmed by empirical evidence, (b) from the two groups of statements the sentence asserting the occurrence of event **E** can be logically deduced" (Hempel 1965:232).

The covering-law model of explanation draws attention to two important characteristics of scientific explanation. First, it provides a logical framework to use in describing explanations: as deductive arguments from general premises and boundary conditions to the explanandum. Second, it places emphasis on the centrality of general laws, laws of nature, lawlike generalizations, etc., in scientific explanation. It thus tries to explain the event in question by showing why it was necessary in the circumstances.

Not all scientific explanations depend on universal generalizations, of course. Some scientific laws are statistical rather than universal. The D-N model has been adapted to cover explanations involving these sorts of laws. The *inductive-statistical* (I-S) model describes a statistical explanation as consisting of one or more statistical generalizations, one or more statements of particular fact, and an inductive argument to the explanandum (Figure

$\mathbf{L}_i$ (one or more universal laws)
$\mathbf{C}_i$ (one or more statements of background circumstances)
_____ (deductively entails)
E (statement of the fact or regularity to be explained)

Fig. 1.2 The deductive-nomological model of explanation

L_i {one or more statistical laws)
C_i (one or more statements of background circumstances)
========= (makes very likely)
E (statement of the fact or regularity to be explained)

Fig. 1.3 Probabilistic explanation

1.3). In this case the form of the argument is different; instead of a deductive argument in which the truth of the premises guarantees the truth of the conclusion (represented by '_____' in the D-N argument), the I-S argument transmits only inductive or probabilistic support to the explanandum (represented by '= = = = ='). That is, it is perfectly possible for the premises to be true and yet the conclusion false. It was noted above that the D-N model interprets scientific explanation of a phenomenon as showing why the phenomenon was necessary in the circumstances. In spite of the formal parallel between the D-N model and the I-S model, they are sharply distinguished because a statistical explanation of an event does not show why it was necessary but rather why it was probable.

Even this account is not sufficient, however. Wesley Salmon shows that many statistical explanations of an event do not even lead to the conclusion that the event was *probable* in the circumstances—only that it was *more* probable in light of the circumstances than it would have been absent those circumstances. Salmon develops his own account of "statistical-relevance" explanations to explicate this feature of probabilistic explanation (Salmon 1984:36 ff.). Suppose, once again, that we are interested in explaining the occurrence of **E** in circumstances **C** and we know various conditional probabilities concerning the occurrence of such events. In particular, we know the probability of **E** occurring in a population **A** (**P**(**E**|**A**)) and the probability of **E** occurring in the subset population satisfying circumstances **C** (**P**(**E**|**A**.**C**)). If we find that P(E|A) ≠ **P(E|A.C),** then C is statistically relevant to the occurrence of **E** (Salmon 1984:32-33); therefore we explain the occurrence of **E** on the basis of the presence of **C**. (We will consider this model again in Chapter 2.)

These represent the main structures that have been offered to represent the logic of scientific explanation. But an adequate account of scientific explanation requires a more substantive discussion. In subsequent chapters we will consider a range of types of explanation—rational choice explanation, causal explanation, structural and functional explanation, and materialist explanation—in substantially greater detail.

Empirical versus theoretical explanation

Social scientists commonly distinguish between empirical and theoretical explanation. The distinction is not well drawn since theoretical explanations, if they are any good, must be empirically supportable. But the contrast is a genuine one that we can characterize more adequately in terms of the distinction between inductive and deductive explanation. An inductive ex-

planation of an event involves subsuming the event under some previously established empirical regularity; a deductive explanation involves deriving a description of the event from a theoretical hypothesis about the processes that brought it about. Suppose we want to know why Bangladesh has a high infant mortality rate. We may seek to explain this circumstance by noting that the nation has a low per capita income (below $200) and that countries with low per capita income almost always have high infant mortality. (As we will discuss in Chapter 8, there is a high negative correlation between infant mortality and per capita income.) In this example we have explained a feature of Bangladesh (high infant mortality) by discovering another feature (low per capita income) with which that feature is usually associated (based on cross-country comparisons).

An important explanatory strategy in science is to attempt to explain a particular phenomenon or regularity on the basis of a *theory* of the underlying structures or mechanisms that produce the explanandum. Theories postulate unobservable mechanisms and structures; for example, physicists explain high-temperature superconductors through a theory of the properties of the exotic ceramics that display this characteristic. Ideally a theory of the underlying mechanisms should permit the derivation of the characteristics of the complex structure; ideally it should also be possible to derive the chemical properties of an atom from its quantum mechanical description.

Consider a typical deductive explanation in social science—a theoretical explanation based on a hypothesis about underlying social mechanisms. Suppose we are interested in the fact that low-level government employees tended to support violent attacks on the state in colonial Vietnam, in contrast to both their better-paid superiors and the less-well-paid, unskilled workers in the city. Why was this particular segment of society stimulated to violent protest? We may try to explain this circumstance in terms of the theory of relative deprivation. This is a theory of individual political motivation that focuses attention on the gap between what an individual expects from life and what he or she is in fact able to achieve. Ted Robert Gurr formulates this theory in terms of the "discrepancy between .. . value expectations and value capabilities" (Gurr 1968:37). Employing this theory we consider the case before us and find that low-level government employees have formed their expectations through comparison with their more privileged colleagues, whereas their incomes are tied to the same economic forces that govern unskilled labor. So when the cost of unskilled labor falls, incomes of low-level government employees fall as well. Finally, we determine that the current economic environment has created a downward pressure on unskilled wages. We now deductively derive a conclusion about the political behavior of low-level government employees: They will be more militant than either high-level government employees or unskilled workers because the expectations of these latter groups match their incomes. Here, then, we have a theoretical explanation of the militancy of low-level government workers.

Both inductive and theoretical approaches to social explanation must confront a particular difficulty. In the case of inductive explanation, we must

ask whether the discovery of a more general empirical regularity embracing the event to be explained is in fact explanatory. Have we arrived at an adequate explanation of Bangladesh's infant mortality rate when we discover the regular relationship between income and infant mortality? It will be argued in Chapter 8 that we need to take a further step and hypothesize the mechanism that connects these variables. In this instance the hypothesis is not difficult to construct: Poor countries and poor families have fewer resources to devote to infant health care, with the predictable result that infants die more frequently. Inductive explanations generally appear to be of intermediate explanatory value. They further our explanatory quest by identifying some of the variables that appear relevant to the event in question. But they should be supplemented by further efforts to provide a theoretical explanation of the empirical regularities that they stipulate.

Turn now to the problems confronting deductive explanation. The central task here is to provide empirical support for the explanatory hypothesis and its application to the particular case. This involves two sorts of investigation: examination of the theory itself in a variety of circumstances and examination of the application of the theory in this particular case. In the relative deprivation case above, then, we must confront several questions. Is it in general true that militant political behavior results from a circumstance of relative deprivation? Further investigation will probably show that the theory describes one of a large number of mechanisms of political motivation: There are instances where individuals' behavior conforms to the theory and other instances where it does not. This does not invalidate the theory, unless the theorist has made rash claims of generality for the theory, but it does mean that we must use care in applying the theory. We must also examine the application of the theory to the particular case. Is there direct evidence showing that low-level government workers define their expectations in terms of the life-styles of high-level government workers? Is there direct evidence showing that their incomes were under stress during the critical period? And is there direct evidence supporting the hypothesis that their militancy was stimulated by this gap between expectation and capability?

Theoretical explanations are essential in social science. At the same time, however, it is important to emphasize the need for careful empirical evaluation of these theoretical hypotheses. What, then, is the function of theoretical analysis in social science? It is to provide the social scientist with an understanding of many of the processes within different social systems—the workings of rational choice, the logic of a market system, the causal influence of norms and values on social behavior, the role of ethnic and religious identity in behavior, and so forth. Social scientists must confront the range of phenomena that constitute their domain with a sensitivity to the diversity of social processes *and* a well-stocked tool box filled with the findings of various parts of social theory.[2]

Nonexplanatory social science

The examples that will be considered in this book have one thing in common: They all represent an attempt to *explain* social phenomena. It

should be noted, however, that explanation is not always the chief goal of scientific research. For example, a common goal of some social research is simply to determine the facts concerning a given social feature. What were the chief characteristics of the Chinese population in the early Qing? Are labor unions effective at increasing safety standards in industry? Was there an industrial revolution? Does U.S. foreign policy ever make use of food as a weapon? In each of these cases, the investigator is primarily concerned with determining an answer to a factual question, one which can only be answered on the basis of extensive analysis and factual inquiry. Clearly, then, there is substantial variety in the forms that social inquiry may take; with its focus on explanation, this book will therefore concentrate on this key aspect of social research.

NOTES

1. This approach parallels that taken by much recent work in the philosophy of psychology. Fodor's work (1980) is a particularly clear example of this stance on the relation between philosophy and an empirical discipline.
2. Stinchcombe (1978) and Merton *(1967)* express this view of the role of theory in social explanation.

SUGGESTIONS FOR FURTHER READING

Achinstein, Peter. 1983. *The Nature of Explanation.*
Braybrooke, David. 1987. *Philosophy of Social Science.*
Elster, Jon. 1983. *Explaining Technical Change.*
Glymour, Clark. 1980. *Theory and Evidence.*
Hempel, Carl. 1966. *Philosophy of Natural Science.*
Miller, Richard *W.* 1987. *Fact and Method.*
Newton-Smith, *W.* H. *1981. The Rationality of Science.*
Rosenberg, Alexander. 1988. *Philosophy of Social Science.*

PART I
MODELS OF EXPLANATION

The following three chapters introduce several central models of social explanation: causal, rational-intentional, and interpretive. These models may be regarded as foundational; they represent the main alternative models of explanation in the social sciences. For a variety of reasons, these approaches are often thought to be in opposition to one another. It is sometimes held that causal explanations are inappropriate in social science because they presume a form of determinism that is not found among social phenomena. Rational choice explanation is sometimes construed as different in kind from causal explanation, and interpretive analysis is sometimes viewed as in-consistent with both rational choice and causal accounts.

It will be argued in this part, however, that such views are mistaken. Causal analysis is legitimate in social science, but it depends upon identifying social mechanisms that work through the actions of individuals. Social causation therefore relies on facts about human agency, which both rational choice theory and interpretive social science aim to identify. It will be held, then, that rational choice theory and (to a lesser extent) interpretive social science provide accounts of the distinctive causal mechanisms that underlie social causation.

2
CAUSAL ANALYSIS

Social scientists are often interested in establishing causal relations among social phenomena—for example, the fact that rising grain prices cause peasant unrest or that changes in technology cause changes in ideology. Moreover social scientists make different sorts of causal claims: singular causal judgments ("the assassination of Archduke Franz Ferdinand caused the outbreak of World War I"), generic causal relations ("famine causes social disorder"), causal relevance claims ("the level of commercialization influences the rate of urbanization"), probabilistic causal claims ("arms races increase the likelihood of war"), etc. Further, a wide variety of factors function as either cause or effect in social analysis: individual actions, collective actions, social structures, state activity, forms of organization, systems of norms and values, cultural modes of representation, social relations, and geographic and ecological features of an environment. (Why, for example, are bandits more common on the periphery of a traditional society than in the core? Because the rugged terrain of peripheral regions makes bandit eradication more difficult.)

The variety of causal claims and variables in social science might suggest that it is impossible to provide a coherent analysis of social causation. But this is unjustified. In fact the central ideas that underlie these various causal claims are fairly simple. This chapter will provide an account of causal explanation within which the variants mentioned above may be understood. And it will emerge that a broad range of social explanations essentially depend on causal reasoning, with certain qualifications. First, the causal assertions that are put forward within social science usually do not depend upon simple generalizations across social properties, that is, they rarely rely on a simple inductive generalization. Second, these claims typically do depend on an analysis of the specific causal mechanisms that connect cause and effect. Third, the mechanisms that social causal explanations postulate generally involve reference to the beliefs and wants, powers and constraints that characterize the individuals whose actions influence the social phenomenon.

THE MEANING OF CAUSAL CLAIMS

What does it mean to say that condition C is a cause of outcome E? The intuitive notion is that the former is involved in bringing about the

latter, given the laws that govern the behavior of the entities and processes that constitute C and E. The social scientist or historian seeks to identify some of the conditions that *produced* the explanandum or that conferred upon it some of its distinctive features. The goal is to discover the conditions existing prior to the event that, given the law-governed regularities among phenomena of this sort, were sufficient to produce this event. There are three central ideas commonly involved in causal reasoning: the idea of a causal mechanism connecting cause and effect, the idea of a correlation between two or more variables, and the idea that one event is a necessary or sufficient condition for another.

In the following, then, I will discuss three causal theses. There is the causal mechanism (CM) thesis:

CM C is a cause of E $=_{df}$ there is a series of events C, leading from C to E, and the transition from each C_i to C_{i+1} is governed by one or more laws L_i.

This definition is intended to capture the idea of a law-governed causal mechanism. Contrast CM with the inductive regularity (IR) thesis:

IR C is a cause of E $=_{df}$ there is a regular association between C-type events and E-type events.

This thesis embodies the inductive model of causation: A statement of causal relation merely summarizes a regularity joining events of type C and events of type **E.** Consider, finally, the necessary and sufficient condition (NSC) thesis:

NSC C is a cause of E $=_{df}$ C is a necessary and/or sufficient condition for the occurrence of E.

This thesis invokes the idea that causes are necessary conditions for the occurrence of their effects and that some set of conditions is sufficient for the occurrence of E.

What are the relations among these conceptions of causation? I will hold that the causal mechanism view is the most fundamental. The fact of a correlation between types of events is evidence of one or more causal mechanisms connecting their appearance. This may be a direct causal mechanism—C directly produces E—or it may be indirect—C and **E** are both the result of a mechanism deriving from some third condition A. Likewise, the fact that C is either a necessary or sufficient condition for **E** is the result of a causal mechanism linking C and E, and a central task of a causal explanation is to discern that causal mechanism and the laws on which it depends.

MECHANISMS AND CAUSAL LAWS

What is a causal mechanism?

I contend that the central idea in causal explanation is that of a causal mechanism leading from C to **E,** so let us begin with that notion. A bolt is left loose on an automobile wheel; after being driven several hundred miles the wheel works loose and falls off. The cause of the accident was the loose bolt, but to establish this finding we must reconstruct the events that conveyed the state of the car from its loose-bolt state to its missing-wheel state. The account might go along these lines: The vibration of the moving wheel caused the loose bolt to fall off completely. This left the wheel less securely attached, leading to increased vibration. The increased vibration caused the remaining bolts to loosen and detach. Once the bolts were completely gone the wheel was released and the accident occurred. Here we have a relatively simple causal story that involves a number of steps, and at each step our task is to show how the state of the system at that point, in the conditions then current, leads to the new state of the system.

Thesis CM above offers a generalization of this mode of explanation. It refers to a series of events connecting C and E. This series of events C_i constitutes the causal mechanism linking C to E, and the laws that govern transitions among the events C_i are the causal laws determining the causal relation between C and E. (In the simplest case the event chain may be very short—e.g., the impact of the hammer produces the smashing of the walnut.) In this account events are causally related if and only if there are causal laws that lead from cause to effect (involving, most likely, a host of other events as well). And we can demonstrate their causal relatedness by uncovering the causal mechanism that connects them.

A causal mechanism, then, is a series of events governed by lawlike regularities that lead from the explanans to the explanandum. Such a chain may be represented as follows: Given the properties of C and the laws that govern such events, C_1 occurred; given the properties of C_1 and the relevant laws, C_2 occurred; . . . and given the properties of C. and the relevant laws, E occurred. Once we have described the causal mechanism linking C to E, moreover, we have demonstrated how the occurrence of C brought about the occurrence of E.

Are there causal mechanisms underlying social phenomena? This question turns in part on the availability of lawlike regularities underlying social phenomena, which will be discussed shortly. Consider a brief example. Suppose it is held that the extension of trolley lines into the outlying districts of a major city caused the quality of public schools in the city to deteriorate. And suppose the mechanism advanced is as follows. Cheap, efficient transportation made outlying districts accessible to jobs in the city. Middle-class workers could then afford to live in the outlying districts that previously were the enclaves of the rich. Over a period of years, an exodus of middle-class workers from the city to the suburbs occurred. One effect of this

movement was the emergence of a greater stratification between city and suburb; prior to suburbanization there was substantial economic mixing in residence, but after suburbanization the poor were concentrated in the city and the middle class in the suburbs. Middle-class people, however, have greater political power than poor people; so as the middle class left the central city, public resources and amenities followed. The resources committed to education in the city fell, with an attendant decline in the quality of public school education in the city.

This story depends on a series of social events: the creation of a new transportation technology, the uncoordinated decisions by large numbers of middle-class people to change their residence, a drop in the effective political demands of the remaining urban population, and a decline in educational quality. Each link in this causal chain is underwritten by a fairly simple theory of individual economic and political behavior, a theory that depends on individuals making rational decisions within the context of a given environment of choice. The story assumes that workers will seek the residences that offer the highest level of comfort consistent with their budget constraints, that they will make demands on local government to expend resources on their interests, and that effective political demand depends a great deal on class. These regularities of human behavior, when applied to the sequence of opportunities described in the story above, led to the changes stipulated. In other words this story describes the mechanism connecting the new trolley system to the degrading of the central city school system.

This example illustrates an important point about causal reasoning in connection with social phenomena: The mechanisms that link cause and effect are typically grounded in the meaningful, intentional behavior of individuals. These mechanisms include the features of rational choice, the operation of norms and values in agents' decisionmaking, the effects of symbolic structures on individuals' behavior, the ways in which social and economic structures constrain individual choice, and so on. This point follows from the circumstance that distinguishes social science from natural science: Social phenomena are constituted by individuals whose behavior is the result of their rational decisionmaking and nonrational psychological processes that sometimes are at work. (Chapter 3 will provide an extensive discussion of the role of rational choice theory in social explanation.)

What sorts of things have causal properties that affect social phenomena? The answers that may be gleaned from actual social explanations are manifold: actions of individuals and groups; features of individual character and motive structure; properties of social structures, institutions, and organizations; moral and ideological properties of groups and communities; new technological opportunities; new cultural developments (e.g., religious systems); characteristics of the natural environment; and more. In each case, however, it is plain how the relevant factor acquires its causal powers through the actions and beliefs of the individuals who embody it.

Consider an example of an explanation that depends on an argument about the mechanisms that mediate social causation (Example 2.1). Kuhn's

Example 2.1 Causes of the Taiping rebellion

There was a pronounced and permanent shift in the balance of power between the Chinese central government and local elites during the mid-nineteenth century. Why did this occur? Philip Kuhn attributes at least part of the explanation to the challenges presented to the Chinese political system by the Taiping rebellion. (a) Elites managed to wrestle control of local militarization from the state bureaucracy and create effective local militias. Prior to the 1840s the state had by and large avoided the use of large local militias to repress banditry and rebellion; after the 1840s it was no longer capable of repressing social *disorder without recourse to* local militias. *(b) Local elites then* effectively managed these organizations against the Taipings. "As the social crisis of mid-century propelled China toward civil war, the pace of local militarization quickened. As economic crises and exploitation drove the poor outside the established order, as scarcity sharpened the conflict among ethnic and linguistic groups, both heterodox and orthodox leadership became increasingly concerned with military organization" (105). (c) Elites managed this because the Qing regime was administratively overextended and because Qing military arrangements were not well designed to control rebellions that increased in scope rapidly. (d) This local militarization ultimately led to a permanent weakening of the center and an increase of local power and autonomy.

Data: historical data on local militia organization in China and the course of the Taiping rebellion

Explanatory model: analysis of local politics and the institutions of the centralized Chinese administration as a basis for explaining the shift in the balance of power between local and national political centers

Source: Philip Kuhn, *Rebellion and Its Enemies in Late Imperial China: Militarization* and Social Structure, 1796-1864 (1980)

---

analysis in Example 2.1 asserts two causal connections: from administrative weakness to the creation of local militias and from the creation of local militias to a further weakening of the political power of the imperial center. Statements (a) and (b) are both factual claims, to be established on the basis of appropriate historical research. But (c) is a claim about the causes of (a) *and (b),* and *(d)* is a claim about *(a)'s* causal consequences. Statement *(c)* represents a "how-possible" question (described in Chapter 1); Kuhn identifies the features of the late Qing administrative system that made it possible for local elites to accomplish what they had not been able to do earlier in the nineteenth century—create local militias that gave them the power to *resist* political imperatives from the center. And *(d)* represents an analysis of the consequences of establishing effective local military organizations: a permanent shift in the balance of power between the state and local elites. The strength of the argument in each case, moreover, turns on the plausibility of the mechanisms through which these changes occurred as specified in Kuhn's historical narrative.

What is a lawlike regularity?

This account of causal mechanisms is based on the idea of a lawlike regularity. Causal relations, then, derive from the laws that govern the behavior of the entities involved. Carl Hempel provides an influential account of causal explanation along these lines in the following passage: *"Causal explanation* is a special type of deductive nomological explanation; for a certain event or set of events can be said to have caused a specified 'effect' only if there are general laws connecting the former with the latter in such a way that, given a description of the antecedent events, the occurrence of the effect can be deduced with the help of the laws" (Hempel 1965:300-301).

A lawlike regularity is a statement of a governing regularity among events, one that stems from the properties or powers of a range of entities and that accounts for the behavior and interactions of these entities. This description, it should be noted, goes beyond interpreting regularities as regular conjunctions of factors; it asserts that they derive from the causal powers of the entities involved. Consider the law of universal gravitation, according to which all material objects attract each other in proportion to their masses and in *inverse* proportion to the square of the distance separating them. This is one of the causal laws that govern the movements of the planets around the sun. The fact that they move in elliptical orbits around the sun is the causal consequence of this law (conjoined with appropriate boundary conditions).

Causal laws may be either deterministic or probabilistic. The law of gravitation is a good example of a deterministic law. All objects, without exception, are governed by this law. (They are subject to other forces as well, of course, so their behavior is not the unique effect of gravitational attraction.) An example of a probabilistic law is Mendel's law of inheritance. If both parents have one-half of a recessive gene—say, blue eyes—then the probability of their offspring having the recessive trait is 25 percent. When the offspring turns up with the trait, we may say that it is the result of the probabilistic law of inheritance of recessive traits; the cause of the outcome is the circumstance that both parents had one-half of the recessive trait.

Are there causal laws among social phenomena? The view that I will defend here is that there are regularities underlying social phenomena that may properly be called "causal," and these regularities reflect facts about *individual* agency. First, the fact that agents are (often and in many circumstances) prudent and calculating about their interests produces a set of regularities encapsulated by rational choice theory—microeconomics, game theory, social choice theory. And second, the fact that human beings conform to a loose set of psychological laws permits us to draw cause-effect relations between a given social environment and a pattern of individual behavior.

Social causation, then, depends on regularities that derive from the properties of individual agents: their intentionality, their rationality, and various features of individual motivational psychology. This finding has

several implications. It follows that social regularities are substantially weaker and more exception-laden than those that underlie natural causation. As a result claims about social causation are more tentative and probabilistic than claims about natural causation. In Chapter 3 I will turn to a discussion of rational choice theory; it is regularities of the sort described there that ultimately provide the ground for causal relations among social phenomena.

It also follows that there are no processes of social causation that are autonomous from regularities of individual action. If we assert that economic crisis causes political instability, this is a causal judgment about social factors. But to support this judgment it is necessary to have some hypothesis about how economic crises lead individuals to act in ways that bring about political instability. (This requirement amounts to the idea that social explanations require microfoundations, a topic discussed in Chapter 9.) When Marx claims, for example, that the institutions of a market economy cause economic crises of overproduction, his argument proceeds through the effects on individual behavior that are produced by those economic institutions and the aggregate effects that individual actions have on the stability of economic institutions over time. This line of reasoning depends on assumptions about what representative economic players do in given circumstances. And these assumptions in turn embody the theory of individual rationality. Institutions and other aspects of social organization acquire their causal powers through their effects on the actions and intentions of the individuals involved in them—and only from those effects. So to affirm that an institution has causal powers with respect to other social entities, it is necessary to consider how typical agents would be led to behave in a way that secures this effect. To say that rumors of bank insolvency are sufficient to produce a run on the bank is to say that, given typical human concerns about financial security and the range of choices available to the typical depositor, it is likely that rumors will lead large numbers of accountholders to withdraw their funds.

THE INDUCTIVE-REGULARITY CRITERION

Let us turn now to the inductive side of causal reasoning (expressed by IR above). Chapter 8 will provide a more extensive discussion of statistical reasoning in social science; in this section we will consider only the basics of inductive reasoning about discrete variables. These are properties that have only a limited number of states—e.g., religious affiliation, marital status, occupation, high-, middle-, and low-income status, etc. This restriction limits us to analysis of causal relations among discrete types of events, individuals, and properties; consideration of correlations among continuous variables must await discussion in Chapter 8. The general idea expressed by IR is the Humean notion that causal relations consist only in patterns of regular association between variables, classes of events, and the like. According to this notion, a pair of variables, **C** and **E**, are causally related if and only if there is a regularity conjoining events of type **C** and events of type **E**. To say that inflation causes civil unrest, in this interpretation, is

	Performance on mathematics test			
Parental income	>90	80-90	70-80	<70
<10,000	.1	1	25	74
10,000-20,000	.1	3	30	67
20,000-30,000	.7	10	35	54
30,000-50,000	5.0	30	40	25
>50,000	5.0	32	43	20

Fig. 2.1 Hypothetical income-mathematical performance data

to say that there is a regular association between periods of inflation and subsequent periods of civil unrest.

The idea of an association between discrete variables **E** and **C** can be expressed in terms of *conditional* probabilities: **E** is associated with **C** if and only if the conditional probability of **E** given **C** is different from the absolute probability of **E**. This condition represents the intended idea that the incidence of **E** varies according to the presence or absence of **C**. Here we are concerned with claims of the following sort: "Marital status is causally relevant to suicide rates." Let **E** be the circumstance of a person's committing suicide and **C** be the property of the person's being divorced. The absolute probability of **E** is the incidence of suicide in the population as a whole; it may be represented as **P**(**E**). The conditional probability of a divorced person committing suicide is the incidence of suicide among divorced persons within the general population; it may be represented as **P**(**E** |**C**) (the probability of **E** occurring given **C**). The statistical relevance test constructed by Wesley Salmon (1984:32-36) may now be introduced: If **P**(**E**) ≠ **P**(**E**|**C**), then we have grounds for asserting that **C** is causally relevant to the occurrence of **E**; if they were *not* causally related, then we should expect that the conditional probability of **E** given **C** should equal the incidence of **E** in the general population. (This is tantamount to the null *hypothesis,* which states that there is no relationship between a pair of values. This idea is considered in greater detail in Chapter 8.)

Suppose we are interested in the causes of the pattern of distribution of superior mathematical ability among high school seniors, as measured by a score of 90 or above on a standard test. Of the total population taking the test only 1 percent falls within the "superior" range, so the absolute probability of a random student qualifying as superior is 1 percent. Now suppose that we break down the population into a series of categories: gender, ethnic background, parental income, parental years of schooling, and student's grade point average. Each classification is designed to be mutually exclusive and exhaustive: Each individual falls within one and only one category. We now produce a series of tables similar to Figure 2.1 for each classification and inspect the conditional probabilities defined by the various cells of the tables, such as the probability of receiving a superior score given a parental income in the $20,000-30,000 range. For some classifications there will not be a significant variation from one cell to

another; each will be approximately 1 percent. In those cases we may judge that the properties defining the classification are not causally related to mathematical performance. In Figure 2.1, however, we find that there is significant variation from one cell to another. The incidence of superior performance among families with incomes of less than $30,000 is significantly below the population average (1 percent), whereas the incidence of superior performance among families with income above $30,000 is significantly above the population average. In other words, mathematical performance is associated with parental income. We may conclude from this finding that parental income is causally relevant to mathematical performance.

This conclusion does not establish the nature of the causal relation. Instead, it is necessary to construct a hypothesis about the causal mechanisms that connect these variables. Several such hypotheses are particularly salient. First, it might be held that superior mathematical capacity is closely related to the quality of mathematical instruction provided to the child and that families with more than $30,000 in income are able to purchase higher-quality instruction for their children. In this case high family income is a cause of high mathematical performance. Second, it might be held that a child's educational experience and the set of cognitive skills that the child develops most fully are highly sensitive to family attitudes toward education, which are in turn correlated with income; families with higher income tend to value education more highly than those with low income. As a result, children of high-income families put more effort into mathematical classwork and on average perform better than children of low-income families. In this case the causal factor is family attitudes toward education, which are (in this hypothesis) tied to income. So income itself is not a causal factor in determining mathematical performance. Finally, it might be held (as Jensen and Herrnstein argued unpersuasively in the 1970s) that performance generally is sensitive to the individual's genetic endowment and that the same genetic features that permitted the parents to attain high income lead to higher-than-average mathematical competence as well. Here we have an instance of collateral causation: Both family income and mathematical performance are effects of a common cause (genetic endowment).

How, then, does the statistical relevance test contribute to an explanation of probabilistic phenomena? Information about conditional probabilities allows us to begin to identify potential causal factors in the occurrence of a characteristic. If one cell of a partition of a population shows a substantially different conditional probability than the base population, the best explanation is that there is a causal factor common to individuals in this cell and not common to the general population that is relevant to the trait in question; otherwise the difference in probabilities can only be the result of random fluctuations that should even out over time. Thus the statistical relevance test supports the inference that there is a causal relationship between E and C. Properly understood, then, the statistical relevance test demands that we back such explanations with some account of the causal factors that give rise to the differing probabilities.

This establishes an important point: Evidence of association gives us reason to believe that there is a causal relationship of some kind affecting the variables under scrutiny, but it does not establish the nature of that relation. Instead, it is necessary to advance a hypothesis about the causal mechanism that produces the observed conditional probabilities. And this hypothesis in turn must be empirically evaluated (perhaps through additional statistical relevance testing [Simon 1971:6]). In the first instance above—wherein the correlation between income and performance is explained as the result of a hypothesized difference in the quality of education provided by higher-income families—it would be possible to design a new study that holds this variable constant and thereby determine whether the conditional probabilities still differ. Our study might compare a significant number of scholarship students from low-income families (with the background assumption that the quality of educational resources will now be equal) and a significant number of high-income students in the same school. If the resulting conditional probabilities are now equal, we have provided empirical support for the educational quality hypothesis. If they are not, then we must consider other hypotheses.

An example of a social explanation that depends explicitly on an inductive method is James Tong's study of collective violence in the Ming dynasty (Example 2.2). Tong's argument may be construed as a "conditional-probability" analysis. The absolute incidence of banditry is .21 events/hundred county-years (the total number of events divided by the total number of county-years embraced by the study). If the variables under scrutiny are causally irrelevant to the occurrence of banditry, then the incidence of banditry in each cell should be approximately .21. The incidence of banditry is broken down into nine cells in Figure 2.2, corresponding to the nine possible combinations of survival risks as peasant and outlaw. In the three cells in the upper right, we find that the incidence of banditry is lower than the absolute incidence for all county-years. In the other six cells, by contrast, the incidence of banditry is greater than the absolute incidence. (This is possible because each cell covers a different number of county-years.) Further there is an orderly progression from the bottom left to the upper right. The highest incidence occurs in the lower left, next come the adjacent cells, and so on toward the upper right cell. There is thus a correlation between the two independent variables and the incidence of banditry. This finding permits us to infer that there is a causal relation between the probability of survival as outlaw and peasant and the occurrence of banditry.

Now we need to identify the causal mechanism that underlies this pattern. Upon inspection it emerges that the cell with the greatest incidence of banditry is the cell in which survival prospects as a peasant are minimum and survival prospects as an outlaw are maximum. But the two cells in which the incidence of banditry is least are those in which survival as peasant is maximum and survival as outlaw is moderate or minimum. Therefore, this finding supports the hypothesis that the occurrence of banditry

Example 2.2 Inductive study of banditry in the Ming dynasty

Banditry and rebellion were common events in imperial China, and they tended to occur in clusters of events across time and space. What caused this temporal and spatial distribution of banditry? James Tong assembles a set of 630 cases of collective violence over the period 1368-1644, distributed over eleven of fifteen Ming provinces. He then evaluates three alternative causal hypotheses:

- Collective violence results from rapid social change;
- Collective violence results from worsening class conflict;
- Collective violence results from situations of survival stress on rational decisionmakers.

He argues that the third hypothesis is correct. He codes each incident in terms of the current "likelihood of surviving hardship" and "likelihood of survival as an outlaw" (Tong 1988:122-24). And he argues that when coded for these variables, the data vindicates the rational choice hypothesis (Figure 2.2).

		Survival as outlaw		
Survival as peasant	Maximum	Moderate	Minimum	Total
Maximum	0.39	0.11	0.12	0.19
Moderate	1.32	0.53	0.20	0.59
Minimum	1.79	0.90	0.82	1.15
Total	0.41	0.13	0.12	0.21

Fig. 2.2 Incidence of banditry per 100 county-years by likelihood of surviving as peasant and surviving as outlaw
Source: Data derived from Tong 1988:126

The most rebellions occur when the probability of surviving hardship is lowest and survival as an outlaw is greatest (1.79 rebellions/county-year), and the fewest occur in the two cells in the upper right (.12 rebellions/county-year). There is a positive association, then, between the variables that Tong isolates and the occurrence of rebellion. Moreover, Tong's causal mechanism to account for this correlation is straightforward; it depends on the rational decisionmaking processes of large numbers of anonymous persons.

Data: a large class of events of social disorder in Ming China culled from local histories

Explanatory model: inductive study used to support the hypothesis that the central causal variable in the occurrence of social disorder (banditry and rebellion) is the rational self-interest of the typical Chinese peasant in changing political and economic circumstances

Source: James Tong, "Rational Outlaws: Rebels and Bandits in the Ming Dynasty, 1368-1644" (1988)

is responsive to the circumstances defining the costs and benefits of banditry for rational agents. When the risks of banditry and the prospects of survival as a peasant are lowest, we should expect that rational agents will be most inclined to adopt the bandit strategy. This expectation is born out in the data produced by Tong.

Let us now evaluate the inductive regularity thesis. It is clear, to start with, that the discovery of an inductive regularity connecting two or more variables strongly suggests a causal relation between the variables. The discovery that electrical workers have substantially higher rates of cancer than the general population is strong evidence that there is some causal influence in their work environment that produces cancers—whether or not we can yet identify the cause. Thus the discovery of regularities, abnormal probability distributions, and correlations is substantial evidence of causal relations. However, the 1R thesis claims more than this; it claims that the notion of a causal relation can be reduced to facts about correlation and conditional probabilities. Is this a defensible claim? It is not because, if applied rigorously, the IR criterion would generate two different sorts of errors (false positives and false negatives). And the best remedy for these failings is to identify the causal mechanisms that produce the observed regularities.

First there is the problem of a spurious correlation between variables (to be discussed at greater length in Chapter 8). Suppose that smokers tend to have nicotine stains on their fingers; that is, there is a correlation between being a smoker and having nicotine stains. If there is a statistical correlation between smoking and cancer, then there will also be a correlation between nicotine stains and cancer. But it is plainly not true that nicotine stains cause cancer. This possibility shows that IR claims too much. The presence of a regularity between two variables does not establish a causal link between them. In this case the IR criterion generates a "false-positive" error: It classifies a relation between two variables as causal when in fact it is not.

The IR criterion may also generate false-negative errors—conclusions that there is no causal relation between two variables when in fact there is. The most prominent source of this kind of error is the possibility of infrequent causal sequences. There may be causal relations among individual events whose covariance is masked when we move to classes of events. In considering a particular rebellion, for example, we may conclude that a famine was the proximate cause of the popular violence, based on an analysis of the particular circumstances and the mechanisms leading from famine to the outbreak of violence. But it may *not* be true that famines and rebellions are correlated; instead rebellions may be greatly dispersed over a variety of background social or economic causes. In this case the IR thesis imposes too coarse a test for causal relations.

To exclude both of these types of errors, we must fall back on an analysis of the possible causal mechanisms that mediate cause and effect. We can best exclude the possibility of a spurious correlation between variables by forming a hypothesis about the mechanisms at work in the circumstances.

If we conclude that there is no possible mechanism linking nicotine stains to lung cancer, then we can also conclude that the observed correlation is spurious. (If we identify the actual causal sequence leading from smoking to both nicotine stains and lung cancer, we can explain the occurrence of the spurious correlation between the latter variables.) Likewise, we can avoid a false-negative error concerning a particular causal sequence (e.g., the occurrence of famine stimulating a particular rebellion) by identifying the causal mechanism that led from one occurrence to the other.

I therefore conclude that the inductive regularity criterion is secondary to the causal mechanism criterion: There is a causal relation between two variables if and only if there is a causal mechanism connecting them. Facts about inductive regularities are useful for identifying possible causal relations, but investigation of underlying causal processes is necessary before we can conclude that a causal relation exists. The 1R criterion should therefore be understood as a source of causal hypotheses and a method to evaluate them empirically—not as a definition of causation.

NECESSARY AND SUFFICIENT CONDITIONS

Causal claims involve identifying necessary and sufficient conditions for the occurrence of an event (principle NSC above). C is causally related to E if and only if C is either necessary for the occurrence of E or sufficient for the occurrence of E (or both). Let us define a *causal field* as the set of conditions that may be causally relevant for the occurrence of the explanandum. A *sufficient* condition C is one in which the presence of C guarantees the occurrence of E. The presence of solar radiation on the dark surface of an object is sufficient to heat the object. The idea that C is a sufficient condition for the occurrence of E corresponds to the intuitive notion that causes *produce* their effects or make the occurrence of their effects unavoidable in the circumstances. However, it is rarely true that any single condition is sufficient for the occurrence of any other. Instead, a group of conditions may be jointly sufficient. So, for example, the material properties of a pane of glass conjoined with the mass and momentum of a baseball are sufficient to cause the window to break. Moreover, causal explanations usually depend on the assumption that "normal conditions" obtain. Suppose that we explain a stock market crash as the effect of investor fears triggered by oil price increases. This explanation requires that we presuppose a set of ceteris paribus conditions: that investors want to maximize gain and minimize losses, that information about commodity prices is available, that investors are free to buy and sell stock, and so forth. But these conditions are part of the normal conditions of a stock market, so they may be taken as fixed. In actual causal arguments in the social sciences, it will often emerge that the claim that C is sufficient for E rests upon an unstated ceteris paribus clause: C is sufficient for E under normal circumstances.

A condition C is said to be *necessary* for the occurrence of an event E if E would not have occurred in the absence of C. The idea that C is a

necessary condition for E reflects the notion that if C is a cause of E, then E would not have occurred if C had not occurred. The presence of oxygen is a necessary condition for the occurrence of combustion; if oxygen is absent, combustion will not occur. Suppose that it is maintained that the assassination of Archduke Franz Ferdinand was a cause of World War I. One way of refuting this claim is to argue that war would have broken out within months even if he had not been assassinated, i.e., the assassination was not necessary for the outbreak of war. This illustrates the phenomenon of *causal overdetermination:* causal fields in which multiple conditions are present, each of which is separately capable of bringing about the event. In such a case none of the circumstances is singly necessary (though it is necessary that one out of a set of circumstances should occur).

We may also distinguish between standing conditions and instigating conditions within a causal field. A standing condition is one that is present over a long period of time and was present for an extended time prior to the occurrence of the explanandum. It is sometimes argued that the naval arms race between Germany and Britain was one of the structural causes of World War I, but this is a condition that extended back to the 1890s. An instigating condition is an event localized in time whose occurrence at time t brought about the occurrence of the effect at time t. An instigating condition introduces the element of change into a state of affairs that produces the effect.

What establishes the relations of necessity and sufficiency among events or conditions? Philosophers have tried to capture these ideas in terms of the concept of natural necessity—the idea that, given the laws of nature and the background circumstances, the former leads unavoidably to the latter.' As we saw above this relation ultimately depends on the causal laws and mechanisms that link cause and effect. Causal laws are the lawlike generalizations—characterizing regularities of human agency, for example—that govern the behavior of the components of the conditions. In the natural sciences, therefore, causal reasoning relies on the assumption that there are laws of nature that establish necessary relations among events and conditions. The claim that the presence of oxygen is a necessary condition for the occurrence of combustion depends finally on our knowledge of the laws of chemistry that govern combustion. Put another way, laws of nature are the basis for our judgment that certain events influence others.

This treatment permits us to construct the following analysis of causal explanation:

A causes **B** if and only if:

1. **A** is a necessary condition for the occurrence of **B;**
2. A belongs to a set of conditions C that are jointly sufficient to give rise to **B.**

However, this account is unsatisfactory for several reasons. First, as we noted earlier, a single condition is almost never a sufficient condition for

the occurrence of another event. Instead the conjunction of a set of conditions is normally needed to supply a sufficient condition; a condition, therefore, may be part of a set of conditions that are *jointly* sufficient for the outcome. The presence of oxygen *and* the presence of dry paper and the presence of a spark are together sufficient for the occurrence of combustion. Thus the presence of dry paper is not sufficient for the occurrence of the fire, nor is it necessary because other combustibles might equally well be present. For reasons of this sort, John Mackie refines the concept of necessary and sufficient conditions by introducing the idea of an INUS condition: an "insufficient but necessary part of a condition which is itself unnecessary but sufficient for the result" (Mackie 1976:62). His point is that there may well be alternative sets of conditions, each of which is sufficient to bring about the event. None of these is necessary because the other sets would do as well. And none of the individual conjuncts of each set is sufficient for the event. Thus Mackie holds that A is a cause of P if and only if it is a part of an INUS condition of P: "A is an INUS condition of a result P if and only if, for some X and for some Y, (AX or Y) is a necessary and sufficient condition of P, but A is not a sufficient condition of P and X is not a sufficient condition of P" (Mackie 1965:237).

The most important defect of the analysis of causal relations in terms of necessary and sufficient conditions is tied to the fact that some causal relations are probabilistic rather than deterministic. Consider the claim that poor communication among superpowers during crisis increases the likelihood of war. This is a probabilistic claim; it identifies a causal variable (poor communication) and asserts that this variable increases the probability of a given outcome (war). It cannot be translated into a claim about the necessary and sufficient conditions for war, however; it is irreducibly probabilistic.

This consideration suggests that the INUS condition is too strong; at best it holds in cases where we have deterministic laws governing the relations among events. But in the case of social phenomena particularly, it is implausible to suppose that the underlying regularities are deterministic. Fortunately, there is an alternative available, in the form of the concept of causal relevance (discussed in the previous section). The concepts of necessary and sufficient conditions can be generalized in terms of comparisons of conditional probabilities. If **C** is a necessary condition for **E**, then the probability of **E** in the absence of **C** is zero (**P**(**E**| —**C**) = 0). If **C** is a sufficient condition for **E**, then the probability of **E** in the presence of **C** is one **(P(E** | **C**) = 1). And we can introduce parallel concepts that are the statistical analogues of necessary and sufficient conditions. **C** is an *enhancing* causal factor just in case **P(E|C)** > **P**(**E**), and **C** is an inhibiting causal factor just in case **P(E|C)** < **P**(**E**). The extreme case of an inhibiting factor is the absence of a necessary condition, and the extreme case of an enhancing causal factor is a sufficient condition.

Consider an example that illustrates a necessary and sufficient condition analysis of social causation (Example 2.3). We may analyze the causal

Example 2.3 Poverty and instability in Latin America

Lars Schoultz analyzes the causal relationship between poverty and instability in Latin America (Schoultz 1987), and his account is summarized in Figure 2.3.

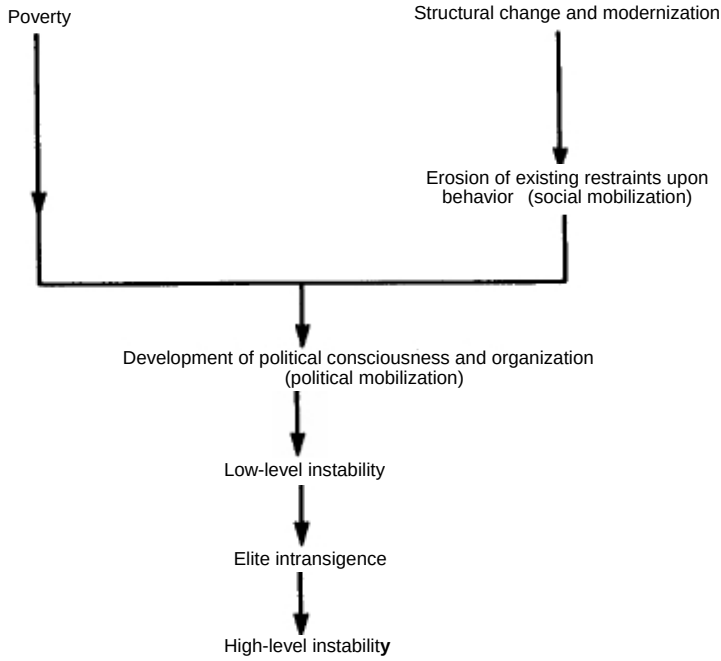

Fig. 2.3 Poverty as a cause of instability

Source: Adapted from Schoultz 1987:72

The arrows in the diagram represent causal mechanisms through which the condition at the top gives rise to the condition at the bottom; thus modernization leads to the erosion of traditional restraints. Schoultz describes this causal hypothesis in these terms: "To be destabilizing, poverty must first await the structural changes that erode traditional restraints upon behavior. Then, when two additional factors—political mobilization and elite intransigence—are also present, the result is instability" (Schoultz 1987:72).

Data: data describing income distribution and political instability in post-1945 Latin America

Explanatory model: causal explanation identifying standing and instigating conditions

Source: Lars Schoultz, *National Security and United States Policy Toward Latin America* **(1987)**

hypothesis in Example 2.3 using the framework we have now developed. Poverty is a standing condition in this analysis, and modernization is an instigating condition. Poverty and modernization are both necessary conditions for the eventual outcome (high-level instability). Poverty, modernization, political mobilization, and elite intransigence are a set of jointly sufficient conditions for high-level instability. Modernization is a historical development that brings its own causal properties—in this case it leads to a weakening of customary or traditional restraints on behavior. For each of these causal processes, we need to provide an account of the mechanisms and laws that give rise to the process. Here modernization erodes customary restraints by disrupting traditional social organization, diminishing the role of the family and traditional religion, stimulating rural-urban migration, and forcing people into market relations. Political mobilization is partly caused by poverty and structural change but not exclusively, which implies that there is an independent unknown causal factor at this point. Once political mobilization has occurred, low-level instability is the unavoidable result. Elite intransigence does not necessarily follow, however; instead we need another independent factor representing the conditions that determine whether elites will be intransigent or accommodating. Finally, if elite intransigence occurs, then high-level instability results through escalating conflict between elites and the poor.

FORMS OF CAUSAL REASONING

In this section I will consider the character of causal reasoning and explore the ways in which social scientists discover or establish causal relations. There are several broad approaches, corresponding to the main elements of the meaning of causal judgments. We will cover comparative analysis and analysis of causal mechanisms here and turn to a more extensive treatment of statistical reasoning in Chapter 8.

The case-study method

Suppose we are interested in explaining the occurrence and character of a particular event—e.g., the Chinese Revolution. Here the research topic may be stated in these terms: Why did the Chinese Revolution occur in the time and circumstances that it did and take the form of a radical peasant revolution rather than an urban liberal democratic movement? This is a causal question. A common approach to such a problem is the *case-study* method, in which the investigator examines the history of the event in detail to arrive at a set of causal hypotheses about its course. The investigator's goal is to discover circumstances in the history of the event that are causally relevant—that is, circumstances that had credible effects on the occurrence, timing, or character of the event. The central difficulty in this type of problem is that we are dealing with a unique series of events, all of which are antecedent to later events in the historical process. Consider three historical circumstances that occurred in China in the 1930s. First, the Great

Depression disrupted the world economy in the 1930s and significantly affected the Chinese rural economy as well. Second, large numbers of Japanese-educated Chinese students returned to China in the 1930s. And third, the Chinese Nationalist movement under Chiang Kai-shek violently expelled the Communist Left from the party in this decade. Each of these circumstances is antecedent to the emergence of a peasant-based political movement aimed at Communist revolution in the late 1930s, and it is possible to interpret each as a causal variable in the occurrence and character of the Communist movement. It might be held that the first and third factors were causally relevant to the Communist revolution but that the second was not. The global depression worsened the economic situation of the peasantry, making that group more easily mobilized by a revolutionary movement. And the Nationalist Party's attack on the Communists impelled the Communist movement to redirect its attention from urban workers to rural peasants. But the return of foreign-educated students had no significant effect on the course of subsequent events. However, this causal analysis must be defended on credible grounds. So it is critically important for the investigator to arrive at a warranted basis for assigning causal importance to diverse factors.

The most common way to support such a causal analysis is by providing an account of the particular causal mechanisms linking various parts of the story. This is one purpose of historical narrative: to establish the series of events that lead from cause to effect. Some links may be non-law-governed—for example, a spontaneous decision by a crucial actor—and others may be governed by social regularities—for example, a price rise in rice relative to wheat leads consumers to shift toward wheat consumption.

To credibly identify causal mechanisms we must employ one of two forms of inference. First, we may use a deductive approach, establishing causal connections between social factors based on a theory of the underlying processes. In this case we note that singular event a is followed by event b, and we argue that this is to be expected on theoretical grounds. Suppose, for example, that it is held that falling prices for cotton in the international market in the 1930s caused Chinese peasant activism. This causal judgment may be supported by a theoretical analysis of peasant political motivation, focusing on the connection between peasant economic security and political behavior.

Second, we may use a broadly inductive approach, justifying the claim that a caused b on the ground that events of type A are commonly associated with events of type B. This reasoning may depend on statistical correlations or on comparative analysis (discussed below). But in either case the strength of the causal assertion depends on the discovery of a regular association between event types.

The construction of a causal story based on a particular case, then, requires two things: fairly detailed knowledge about the sequence of events within the large historical process and credible theoretical or inductive hypotheses about various kinds of social causation. Consider the hypothesis

that the depression increased the likelihood that a revolutionary peasant movement would succeed. This hypothesis depends on several kinds of knowledge. It presupposes a theory of political behavior: Peasants are concerned about their economic welfare, and the worse their economic circumstances, the more likely they are to support radical political movements. It also requires fairly detailed historical knowledge about the rural economy and peasant political behavior in the 1930s: We need to know whether the rural economy did in fact worsen during the 1930s and whether peasants did in fact become more responsive to radical movements as conditions worsened. If these assumptions are not born out, then the causal hypothesis fails. Finally, this causal argument is much strengthened if it can be supported with comparative and inductive evidence. If the researcher can show that radical political movements in other settings (Vietnam, Cuba, Ming China) have been sensitive to worsening economic circumstances, this provides empirical support for the singular causal judgment in this case as well.

These considerations lead us to some conclusions about the case-study method. It involves the detailed study of a particular sequence of social events and processes. And it depends on identifying particular causal links among historical events and circumstances. But the claim of causal connectedness unavoidably requires more than the knowledge of temporal succession among the events; we also need a theoretical or inductive basis for asserting that a given historical circumstance affected the occurrence and character of a subsequent circumstance. This leads us, then, to several other forms of causal reasoning, especially the comparative method and analysis of particular causal mechanisms.

Consider an example of a case-study analysis of social causation—Elizabeth Perry's explanation of the Nian rebellion in North China (Example 2.4). Perry's analysis is based on a detailed study of one extended historical event—a major peasant rebellion. And she arrives at a hypothesis about the conditions that caused this event: a set of environmental and social circumstances that provided individuals with an incentive to support bandit and rebel organizations in order to survive. Finally, her account depends on the theoretical analysis of individual decisionmaking within a particular environment of choice.

The comparative method

Another important approach to causal analysis is the comparative study of cases that embody a range of similar characteristics with certain salient differences. What explains different outcomes in apparently similar circumstances? For example, why do some poor villages become more cohesive in the face of famine, war, or flood and others become less so? Are there general factors that account for these differences? Or are the differences the result of historical accident?

In the comparative approach the investigator identifies a small number of cases in which the phenomenon of interest occurs in varying degrees and then attempts to isolate the causal processes that lead to different

Example 2.4 Peasant rebellion and strategies of survival

Peasant rebellions were a recurring feature of nineteenth-century China. What caused these rebellions? Elizabeth Perry analyzes the Nian rebellion that occurred in North China in the 1850s. After detailing the precarious ecology of North China, Perry holds that the central concern of peasants in this area was to find and pursue a strategy of survival. She identifies two broad families of such strategies: predatory and protective. Predatory strategies include salt-smuggling, petty theft, and banditry; protective strategies involve largely village-level defense organizations (militias, fortification, etc.). She argues that the Nian rebellion was the unintended outcome of the interaction between these strategies: As bandit gangs became more attractive to desperate peasants, bandit predations became more dangerous to villages and conflict escalated between militias and bandit gangs. Eventually bandit gangs grew large enough to attract the attention of the state, and in self-defense they organized themselves to repel military attack by state forces. Thus Perry holds that the Nian rebellion should be understood on the basis of factors at the level of the peasant household and village, not national or regional political factors. And she pays close attention to the circumstances at the local level that made it rational for individual peasants either to support local militias or to join bandit gangs.

Data: nineteenth- and twentieth-century peasant political behavior in the North China plain

Explanatory model: rebellion was the aggregate result of individually rational strategies of survival that escalated to large-scale collective action

Source: Elizabeth Perry, Rebels and *Revolutionaries in North* China 1845-1945 (1980)

---

outcomes. This method requires a close scrutiny of the details of the cases, along with an effort to develop a hypothesis about the cases' causal dynamics. Thus comparative studies look at the details of a few cases in order to probe the mechanisms of change, the details of the processes, and the presence or absence of specific factors. The comparative study often uses a form of Mill's methods (discussed below), reasoning that if a given outcome is present in one case and absent in the other, there must be a causal factor present in the first case that is lacking in the latter. And the comparative method looks directly for causal mechanisms through which differing out-comes result from given social circumstances.

Theda Skocpol is a prominent exponent of the comparative method for social science. She describes her method in these terms: "The overriding intent is to develop, test, and refine causal, explanatory hypotheses about events or structures integral to macro-units such as nation-states" (Skocpol 1979:36). The comparative method is applied to a fairly small number of cases involving large social units in which the explanandum phenomena are found. The method then proceeds by identifying a set of relevantly similar cases involving the phenomenon to be explained—in Skocpol's case, the occurrence of successful revolution in France, Russia, and China. As

Charles Ragin describes it, "Comparativists are interested in the similarities and differences across macrosocial units" (Ragin 1987:6). The investigator then tries to determine whether there are factors that covary across the cases in such a way that they can be potential causes of the phenomenon to be explained.

Consider a hypothetical example. Suppose we are concerned with the occurrence of popular social conflict—riots, eat-ins, rebellions, etc. Using the comparative method we would identify several cases in which there is a substantial history of such conflict—say colonial Vietnam, seventeenth-century France, and Qing China. We would first pursue a detailed under-standing of the processes of social conflict in each of the cases. Then we would try to determine whether there are similar patterns in the several cases. Now suppose that it is suggested that sharp class conflicts are a necessary and sufficient condition for the occurrence of social conflict. A comparative study can do two things. It can determine that revolutions have occurred in the absence of class conflict—thus refuting the claim that class conflict is a necessary condition for revolution. And it can determine that there are circumstances in which there was intense class conflict but no revolution—thus refuting the claim that class conflict is a sufficient condition.

Suppose that we find that class conflict was present in all the positive cases and absent in the negative ones, i.e., that class conflict covaries with revolution exactly (which it does not, in fact). Does this establish that class conflict is a necessary and sufficient condition for the occurrence of revolution? It does not, for two reasons. First it is possible that the covariance is accidental or artifactual; whenever we are restricted to an examination of a small number of cases, it is always possible that covariance is the result of random events. And second we have the familiar problem of spurious correlation: It may be that both class conflict and successful revolution are the collateral effects of some third factor. To exclude these possibilities we must construct a theory of the mechanism connecting cause and effect—the pathway by which the explanans gives rise to the explanandum.

Theda Skocpol's analysis of the causal conditions of successful revolution represents an important instance of comparative analysis (Example 2.5). Skocpol's analysis treats social unrest as a standing condition that is present in virtually all agrarian societies. Therefore, she suggests, social unrest cannot be the immediate cause of revolution—otherwise all agrarian societies would undergo revolution. It is therefore necessary to find a factor that is present in the instances in which revolution occurs and absent otherwise. And Skocpol argues that the factors that vary in the appropriate way are the competence and coherence of the state and its capacity to preserve itself in the face of popular opposition. Note, however, that this argument does not demonstrate that social tension is not a causal factor in the occurrence of revolution, only that it is not a sufficient condition. On this account, social tension is a necessary condition for the occurrence of revolution, and, when it is experienced in a society characterized by a weak state, revolution ensues.

Example 2.5 State structure and revolution

What explains the success of revolutions in a small number of cases and the failure of revolutionary movements in many others? Theda Skocpol offers a comparativist analysis of the causes of revolution in China, France, and Russia to answer this question. She argues for a complex causal hypothesis: that peasant unrest is a necessary but not sufficient condition for social revolution in pre-industrial societies, that such unrest is virtually ubiquitous and that the critical variable determining whether revolution occurs is the status of the state structure. Her causal account therefore focuses on the administrative capacity and competence of the state. She holds that the three revolutions studied all showed the same pattern: Old regime states were confronted with international crises they could not handle, and in those circumstances endemic class conflicts broke out that the repressive and political powers of the state were incapable of eliminating. She writes, "| have argued that (1) state organizations are susceptible to administrative and military collapse when subjected to intensified pressures from more developed countries abroad and (2) agrarian sociopolitical structures that facilitated widespread peasant revolts against landlords were, taken together, the sufficient distinctive causes of social-revolutionary situations commencing in France, 1789, Russia, 1917, and China, 1911" (Skocpol 1979:154). |n this account the critical factors that determined whether rebellion would occur were the structure of the state and the social and political arrangements that governed local life.

Data: comparative study of the social, economic, and political circumstances that preceded the French, Russian, and Chinese revolutions

Explanatory model: a structural-causation model, according to which variations in the political structures of several societies account for the success or failure of revolution in those societies

Source: Theda Skocpol, States and Social Revolutions: A Comparative *Analysis of* France, Russia, and China (1979)

Consider a second example of comparative analysis: Atul Kohli's analysis of the politics of poverty reform in India (Example 2.6). Kohli's analysis begins by identifying the factor to be explained across cases—the existence and effectiveness of poverty-alleviation programs. He then attempts to determine the features of social and political institutions that covary with this factor and plausibly represent the primary causal mechanisms that account for differences in the factor. His account presupposes a specification of the causal field—that is, the factors that are potential causal variables prior to investigation. (So, for instance, Kohli does not consider ethnic composition as a potential causal variable.) Finally, he argues that there is a complex political factor whose presence or absence covaries in the predicted way with the existence and effectiveness of poverty programs—the political ideology and competence of the regime in power. He concludes that this factor is the primary causal variable in producing the different outcomes. This argument, it should be noted, proceeds both inductively and deductively.

Example 2.6 Poverty reform in India

Atul Kohli notes that the situation of the poor in India has scarcely changed *since* independence *in 1947, in* spite *of the economy's* respectable *rate of* growth in that period. However some states in India have done better than others in poverty alleviation. What are the social and political factors that influence the welfare of the poor in the process of third-world economic development? Kohli undertakes a comparative study of the economic policies of three Indian states (West Bengal, Karnataka, and Uttar Pradesh). He finds that the welfare of the poor is not correlated with the overall prosperity of a state. Instead, the critical variable is the type of regime in power during the process of economic development. Regimes formed by strong, competent political parties of the Left succeed in tilting the process of development toward poverty
alleviation, whereas weak regimes and those dominated by the propertied classes have a poor record of performance in poverty reform. The Communist Party, Marxist (CPM) in West Bengal succeeded in bringing tangible benefits to the poor through poverty reforms including tenancy reform and rural credit and employment programs. CPM is a leftist party with a coherent redistributivist ideology, competent party organization extending down to the village level, and effective leadership. The Urs regime in Karnataka also possessed a redistributivist ideology but lacked effective political organization and had a fragmented leadership; its efforts at poverty reform were not successful. And the Janata Party in Uttar Pradesh was dominated by the rural landowning class and lacked the will to implement poverty reforms. Kohli explains the presence or absence of poverty alleviation in a state, then, as the result of the presence or absence of a regime that has both the will and the means to implement poverty reform.

Data: *economic* and political data drawn *from* three Indian *states in the 1970s*
Explanatory model: a causal explanation of poverty reform in India based on comparative analysis of the political aims and capacities of different regimes and parties
Source: Atul Kohli, *The State and Poverty in India: The Politics of Reform* (1987)

The inductive side corresponds to the point about covariance between regime type and poverty performance, but the deductive side takes the form of a theoretical argument designed to show why this result is a plausible one. In other words, Kohli's position relies on an argument about the causal mechanisms through which poverty policies are adopted and implemented in state governments in India.

Mill's methods

The comparative method depends heavily on an analysis of causal reasoning provided by John Stuart Mill in his *System of Logic:* the methods of agreement and difference. These are methods aimed at identifying the cause of an event by observing variations in antecedent conditions for repeated occurrences of the event.[2] Suppose that we are interested in discovering the cause of an event P in a causal field of a range of possibly

	P	A	B	C	D	E
I_1	p	p	p	a	a	p
N_1	p	p	a	p	a	a

Fig. 2.4 Mill's method of agreement

	P	A	B	C	D	E
I_1	p	p	p	a	a	p
N_1	a	a	p	a	a	p

Fig. 2.5 Mill's method of difference

relevant factors {**A,B,C,D,E}.** For vividness, suppose that the event **P** is the success of a union-organizing drive and the causal factors are: (**A**) falling real wages, (**B**) urban setting, (**C**) skilled labor force, (**D**) authoritarian management style, and (**E**) industrial company. That is, we are interested in discovering a factor that is necessary and sufficient for the occurrence of **P**. The method of agreement instructs us to find two or more cases in which **P** occurs and in which only one of the possible causal factors is present in all cases (factor A in Figure 2.4). (The letters **p** and **a** signify the presence or absence of the factor in question.) In this example, then, we need to find two or more instances of union-organization drives that lead to success and then determine the state of factors **A** through **E**. If the set of factors surveyed is exhaustive and if there is a single necessary and sufficient condition for the occurrence of **P**, then the factor that is present in every case must be the necessary and sufficient condition. Here it is the "real wage" variable that is constant across the cases, so the method of agreement would lead us to conclude that the direction of change of real wages is the cause of success or failure in union-organizing drives.

Turn now to the method of difference. In this instance we are instructed to find a pair of cases in the first of which **P** occurs and in the second of which it is a sent. Once again we are to survey the set of relevant factors

{A,B,C,D,E}. If there is a single factor that covaries with **P**, we can conclude that **A** is the cause of **P**. In Figure 2.5 there are two cases, one in which **P** occurs ($\mathbf{I}_1$) and one in which **P** does not occur (**N**,). We now survey the two circumstances and find that **B, C**, **D**, and **E** remain fixed through both cases, and **P** and **A** vary from the first case to the second. We can conclude from this analysis that **C** and **D** are not necessary conditions for **P** because they are absent in $\mathbf{I}_1$ The only factor that is present when and only when **P** occurs is **A**. If **B** were a sufficient condition for the occurrence of $\mathbf{P}_1$ then **P** ought to have occurred in $\mathbf{N}_1$ as well. Therefore, the method of difference permits us to conclude that **B** is not a sufficient condition for the occurrence of **P**.

But do these findings permit us to conclude that **A** is a sufficient condition for P? They do so only if we can assume that {**A,B,C,D,E**} is an exhaustive set of causal factors for the occurrence of P; otherwise it is entirely possible

that the covariance of **A** and **P** is accidental. However this is a highly unrealistic assumption; in the typical case it will be an open question whether there are other as yet unidentified causal factors. If we do not know that **{A,B,C,D,E}** is exhaustive, t en the best we can conclude is that only **A** out of the set **{A,B,C,D,E}** is potentially a necessary and sufficient cause of **P** and only A, **B,** and **E** are potentially necessary conditions for **P**. To have further reason to suppose that **A** is sufficient and necessary, we need to survey a number of other possible cases. Ideally it will emerge that **A** always covaries with **P**, and neither **B**, **C**, **D**, nor **E** is necessary for the occurrence of **P**.

What Mill's methods cannot handle are complex causation and probabilistic causation. Suppose that **A** causes **P** when in the presence of **F** and **B** causes **P** when in the presence of **G**. Then there will be cases where **A** is absent, **B** is present, and **P** is present; there will be cases where **A** is present, **B** is absent, and **P** is present; and there will be cases where **A**, **B,** and **P** are all present. The first such case would indicate that **A** is not a cause of **P**, and the second indicates that **B** is not a cause of **P**. Likewise suppose that **A** is the only cause of **P**, but it is a probabilistic cause: If **A** occurs, then there is a 90 percent chance that **P** will occur as well. If our set of cases includes one of the rare instances where **A** occurs and **P** does not, the method of difference will exclude **A** as a cause of **P**. Thus Mill's methods are well designed only for cases where we have single conditions that are necessary and sufficient for the occurrence of the outcome. Moreover, these methods require relatively demanding conditions for their application: a complete list of potentially relevant causal conditions, a pair of observations in which **P** occurs and does not occur, and information about the occurrence or nonoccurrence of each of the relevant conditions. In spite of these limitations, however, Mill's methods underlie much reasoning about causation in the social sciences.

CONCLUSION

The fundamental idea underlying causal reasoning in social science is that of a causal mechanism: To claim that **C** caused **E** is to claim that there is a causal mechanism leading from the occurrence of **C** to the occurrence of **E**. We have seen that this concept is the basis for two other prominent ideas about causation: the ideas that causal judgments correspond to inductive regularities and express claims about necessary and sufficient conditions. We have also seen that the discovery of an inductive regularity between two variables is a strong reason to expect a causal connection between them, although the connection itself takes the form of a causal mechanism. Likewise if it is true that there is a causal mechanism connecting **C** and **E**, then it follows that the occurrence of **C** enhances the probability of the occurrence of **E** (the most general version of the necessary and sufficient condition thesis).

Subsequent chapters will show that causal explanation plays a very prominent role in social science. We will find that materialist, functionalist,

and structuralist explanations may be seen as specialized forms of causal explanations. And it will emerge that statistical explanations in social science, when they are genuinely explanatory, depend on the availability of credible hypotheses on underlying causal mechanisms. It is commonly held that there are distinctive noncausal explanations available to the social sciences—for example, structuralist, rational-intentional, or interpretive explanations. But arguments in later chapters will cast doubt on this view. We will show that the central causal process underlying social change derives from rational-intentional behavior on the part of individuals. Thus there is an intimate connection between causal and rational explanation, which will be explored in the next chapter. The sole exception to the idea that social explanations are primarily causal explanations is the interpretive social science paradigm—a framework that we will consider in Chapter 4. And Chapter 5 will show that functional and structural explanations, when valid, are specialized forms of causal explanations.

NOTES

Rom Harre (1970) develops this view in detail. Related views may be found in Salmon (1984).

Mill's methods are described in Mill (1950). Discussion of the methods can be found in Mackie (1974:68 ff.).

SUGGESTIONS FOR FURTHER READING

Elster, Jon. 1983. Explaining *Technical* Change.
Mackie, J. L. 1974. Cement *of* the Universe.
Miller, Richard W. 1987. *Fact and Method*.
Ragin, Charles C. 1987. *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*.
Salmon, Wesley C. 1984. *Scientific Explanation and the Causal Structure of the World.*
Skyrms, Brian. 1980. *Causal Necessity: A Pragmatic Investigation of the Necessity of Laws.*

3
RATIONAL CHOICE THEORY

Social phenomena result from the activities of human beings, and human beings are *agents* whose actions are directed by their beliefs, goals, meanings, values, prohibitions, and scruples. Human beings, that is, are *intentional* creatures who act on the basis of reasons. This has a number of implications for the social sciences. First, it implies that social regularities derive from a rather different type of causal relation than do natural regularities. The latter stem from the fixed, objective features of the entities involved and the laws of nature that govern them, while the former stem from the intentional states of the agents. Second, the intentional character of social phenomena makes possible a type of explanation for social science that is not available in natural science. Many social phenomena can be explained as the aggregate consequence of the purposive actions of a large number of individuals. By coming to understand what those persons wanted, what they believed, and how they expected their actions to further their goals, we can explain the occurrence of the aggregate consequence as well.

In this chapter we will examine a model of explanation based on this feature of social life—aggregative explanations that attempt to account for social patterns as the aggregate result of the rational actions performed by large numbers of participants. Rational choice theory provides a formal analysis of rational decisionmaking on the basis of a set of beliefs and goals, and it incorporates several areas of economic theory—probability theory, game theory, and the theory of public goods. In the previous chapter we found that causal explanations of social science require some account of the mechanisms that mediate between cause and effect. The rational choice paradigm offers a general account of such mechanisms among social phenomena. If we can assume that individuals in a variety of social settings make calculating choices based on their beliefs and goals, we may be able to explain numerous social arrangements as the aggregate effect of such choices. This paradigm is controversial, however, for some social scientists believe that the rational choice approach abstracts too much that is culturally specific in human action, with the result that rational choice "theorems" have little to do with actual social behavior. This chapter presents some of the fundamental ideas of the rational choice paradigm. And in later chapters

Example 3.1 Feudal labor services and economic rationality

European feudalism was characterized by a legal obligation of the peasant to provide labor services for the lord. This is one system of surplus extraction, but there are many others—fixed wages, fixed rents, or some combination. Why were compulsory labor services selected by the manorial economy as the form of surplus transfer from peasant to lord? Douglass North and Robert Paul Thomas interpret feudalism as an exchange of goods between lord and peasant; the lord provides various public goods—chiefly security—and the peasant provides part of his surplus as income to the lord. North and Thomas argue that the labor service contract is the most acceptable arrangement to both lord and peasant in the context of a nonmarket economy. Fixed wages require the lord to assume the risks of cultivation (because wages must be paid whether the crop is successful or not), and fixed rents require the peasant to assume the risks; in either case the costs of negotiation between lord and peasant are high because the necessities of life are difficult to evaluate in the absence of a monetized economy. A labor service arrangement, on the other hand, provides a standard arrangement that is *easy* to negotiate and enforce and automatically adjusts to both good and bad years. "The contractual arrangement of the classic manor can now be seen as an efficient arrangement for its day. The obligation of the serf to provide labor services to his lord and protector, an input-sharing arrangement, was chosen because given the constraint of high transaction costs involved in trading goods it was the most efficient. . . . The `quaint' organization of the classic manor is therefore understandable as an appropriate response in the general absence of a market economy" (North and Thomas 1973:31-32).

Data: historical data about the manorial economy and the legal relations between lord and tenant

Explanatory model: explain patterns of human behavior as the outcome of deliberation within the framework of economic rationality

Source: Douglass C. North and Robert Paul Thomas, The Rise of the Western *World:* A New Economic History (1973)

---

we will see how these ideas are applied to concrete problems of social explanation in economic anthropology, public choice theory, and Marxist theory.

Example 3.1 illustrates the aggregative mode of explanation. Here North and Thomas explain the system of bonded labor as the most advantageous to both serfs and lords; they hold that the labor service contract was selected by participants within feudal society because it was the most economically efficient arrangement available and was in the interest of both lord and peasant. On this account, then, a key feature of feudalism is explained as the aggregate consequence of the rational choices made by large numbers of peasants and lords over time.

AGGREGATIVE EXPLANATION

The rational choice paradigm of explanation rests on one central premise and a large set of analytical techniques. The premise is that individual

behavior is goal-directed and calculating. Individuals are assumed to have a set of interests against which they evaluate alternative courses of action; they assign costs and benefits to various possible choices and choose an action after surveying the pros and cons of each. Rational choice explanations thus depend upon the "means-end" theory of rational action. An action is rational just in case it is an appropriate means of accomplishing a certain end, given one's beliefs about the circumstances of choice. Therefore, to explain an individual's action is to identify his or her background beliefs and goals and to show how the action chosen is a reasonable way to achieve those goals given those beliefs.'

This account of rationality may be described as a "thin" theory of human action.[2] It depends on an abstract description of goals in terms of interests, utilities, or preferences and postulates a simple mode of reasoning—utility maximization, for example. On the basis of these simplifying assumptions, rational choice theorists hope to explain a variety of human behaviors. The advantage of this approach is explanatory parsimony and power; to the extent that these assumptions bear some relation to human behavior, they provide the basis for explaining a wide range of social phenomena in a variety of cultural settings. However, a primary criticism of rational choice analysis arises at this point because interpretive social scientists postulate the need for "thick" descriptions of human action—detailed accounts of norms and values, cultural assumptions, metaphors, religious beliefs and practices—in order to account for human behavior. Furthermore, they deny that more abstract descriptions of human action are of much explanatory value. We will return to these criticisms in the next chapter.

So far we have not considered the content of the goals that guide individuals' actions. Economists, however, tend to include at least one substantive assumption in their account of rationality—the assumption of egoism. They assume that each economic agent is solely concerned with maximizing his own *private* interests—minimizing labor, maximizing income, maximizing leisure, and so forth. However this assumption is not essential to rational choice theory; it is possible to leave open the question of the nature of the agent's goals. In this light, the problem of rational choice theory is how to specify the best way of deciding among a range of choices given one's ends. The content of the agent's ends is left open; some individuals may attach utility to self-interest, the interests of various other persons, and the public good, while others may be solely concerned with self-interest.

A final issue raised by the thin conception concerns the rationality of beliefs about the environment of choice. This factor reflects the fact that rational action depends on the agent's possessing beliefs about (1) the options that are available to him or her and (2) the probable consequences of each action. This presents us with a choice in formulating a thin theory of rationality: Shall we require that the agent's beliefs about the probable consequences of the outcomes are themselves rationally grounded—that is, shall we require that the agent has rational beliefs—or shall we take the agent's beliefs as given and focus only on the problem of choice relative to those beliefs? I will assume that the thin theory involves both rational

beliefs and rational choices, so that I will also assume that rational agents come to their beliefs about the consequences of their actions on the basis of appropriate inductive methods.

How does the concept of individual rationality give rise to explanations of social phenomena—the occurrence of collective action, enduring social institutions, or processes of social and economic change? The rational choice approach seeks to explain social outcomes as the aggregate result of large numbers of individuals acting on the basis of rational calculations. Malthus's predictions about the relation between economic trends and population curves depends on this assumption, as do Marx's analysis of the capitalist economic system and contemporary "political economy" approaches to politics in peasant societies. What these theories have in common is an explanatory strategy: explaining a social pattern as the aggregate consequence of the rational actions of a large number of participants, given the circumstances of the social and natural environment within which they deliberate. Why do strikes often collapse before they gain their objectives? Because defection has advantages for individual strikers. Why do prices tend to oscillate around the cost of production plus an average rate of profit? Because rational entrepreneurs enter and exit industries according to the rate of profit. Why do arms agreements tend to break down? Because participants fear unilateral defection by their opponents. Thus Elizabeth Perry explains the emergence of Nian armies as the aggregate result of local predatorial strategies of survival (Example 2.4); Samuel Popkin explains the failure of collective action in village societies as the effect of free-rider choices (Example 7.1); and Robert Brenner explains the stagnation of French agriculture as the absence of incentives and opportunities for technological innovation on the part of landlords and peasants (Example 6.6). In each case the author identifies a pattern of rational individual behavior that responds to a particular set of incentives and constraints and then attempts to show how this pattern of individual behavior aggregates into the observed macropattern.

These efforts may be described as *aggregative explanations,* which seek to explain large-scale social, economic, and political phenomena as the aggregate and often unintended outcome of rational decisionmaking at the individual level, Here the formal tools of rational choice theory are of value for they offer a variety of analytical techniques for deriving the aggregate effects of the actions of a large number of rational decisionmakers. Game theory, collective action theory, and marginalist economic theory each provide aggregation techniques for a range of situations within which rational decisionmakers act: strategic conflict and cooperation, public goods problems, and markets. The motivational and systemic conditions defined by social institutions impose discernible patterns on society in this sense: They define both the interests that guide various actors within society and the prohibitions and incentives that influence deliberation. They thus represent a highly structured system within which individuals act, and they impose a pattern of development and organization on society as a whole. Explanation therefore consists of showing the process through which these conditions shape the

Example 3.2 Residential segregation

Noting the common pattern of segregation between ethnic groups in U.S. cities, Thomas Schelling attempts to construct an explanation of this in terms of a hypothesis about the preferences of individuals. "This chapter is about the kind of segregation ... that can result from discriminatory individual behavior. ... It examines some of the *individual* incentives and individual perceptions *of* difference that can lead *collectively* to segregation" (Schelling 1978:138). *He* shows that rather weak assumptions about individual preferences are sufficient to produce sharply segregated residential patterns in the aggregate. In particular, if we assume that members of each ethnic group will tolerate an ethnically mixed neighborhood up to a certain ratio and will move if the proportion rises above that ratio, in a variety of neighborhood models it emerges that the stable equilibria are those in which the two groups are sharply segregated. This aggregate result stems not from the fact that each person prefers to live in a segregated neighborhood but rather from the ripple effects that follow as residents in unsatisfactory neighborhoods move into new neighborhoods, thereby altering the proportions in the new neighborhood and stimulating new movement.

Data: descriptive data concerning residential patterns in a variety of cities in the world

Explanatory model: aggregative explanation based on a hypothesis about agents' neighborhood preferences

Source: Thomas Schelling, *Micromotives and Macrobehavior* (1978)

observable features of the social system. Examples 3.2 and 3.3 illustrate this mode of explanation.

Schelling's explanation (Example 3.2) is a simple one. On the basis of an uncomplicated hypothesis about individual preferences, he derives the aggregate consequence of those preferences within a simple model. Marx's model in Example 3.3 is slightly more complex but essentially similar. It can be summarized in the following way: A given feature of capitalism occurs because capitalists are rational and are subject to a particular set of incentives, prohibitions, and opportunities. When they pursue the optimal individual strategies corresponding to these incentives, prohibitions, and opportunities, the explanans emerges as the aggregate consequence of the resulting choices. Each of these is thus an aggregative explanation because it attempts to show that a social feature is the unintended consequence of the rational strategies chosen by large numbers of participants within a particular environment of choice.

The rational choice approach, then, rests upon a simple explanatory strategy. To explain a given social phenomenon it is necessary and sufficient to provide an account of:

- the circumstances of choice that constitute the environment of action;

Example 3.3 Marx's economics

Nineteenth-century capitalism displayed a number of systemic characteristics—for example, crisis, concentration of capital, a falling rate of profit, and a pool of chronically unemployed workers. Marx sought to explain these characteristics (which he called "the laws of motion of the capitalist mode of production") through analysis of the defining economic institutions of capitalism: production for profit organized around independent, privately owned, labor-hiring firms. The capitalist economy is defined by a set of social relations of production (property relations). These relations determine relatively clear circumstances of choice for the various representative actors (the capitalist, the worker, the financier). And these circumstances are both motivational and conditioning: They establish each party's interests, the opportunities available to each, and the constraints on action that limit choice. The problem confronting Marx is that he must demonstrate, for a given characteristic of the capitalist mode of production (e.g., the falling rate of profit), that this characteristic follows from his account of the primary institutions of capitalism through reasoning about rational behavior within the circumstances of choice. Capitalists strive to maximize the rate of profit in their firms, which leads them to adopt cost-cutting new technologies that are typically capital intensive; when these innovations are adopted by all producers, the rate of profit falls.

Data: economic indicators of nineteenth-century capitalism (rate of profit, size of firm, wage data, etc.)

Explanatory model: aggregative explanation based on (1) rational individual capitalist behavior, (2) the constraints and incentives created by the capitalist economic structure, and (3) use of classical economic models to derive consequences from these findings

Source: Karl Marx, Capita!, vol. 1 (1867/1977)

- the strategies that rational, prudent persons would pursue in those circumstances;
- the aggregate effects of those strategies.

Social phenomena, in this view, are the result—often unintended—of the purposive actions of large numbers of rational agents, and explanation consists in showing how the circumstances of individual action stimulate the patterns of behavior that in turn give rise to the observed social phenomena.

This model requires further analysis at two points. First, we need a formal account of the structure of rational decisionmaking so that we can arrive at determinate predictions about rational choice in particular social circumstances. Second, we need an analysis of some of the situations of interactive social behavior to which the rational choice approach may be applied, specifically strategic rationality and collective action. The following sections will consider each of these aspects of the aggregative model of explanation.

DECISION THEORY

This section will offer a closer examination of the details of the rational choice framework. I will also discuss the foundations of rational choice theory—the notions of utility, probability, and a decision rule.

Utility and preference

The thin theory of rationality may be stated as follows: "Agents act rationally insofar as they choose their actions from the range of available *options* that best serve their ends, given their *beliefs* about available options and their probable consequences." The thin theory assumes that agents have a consistent set of aims or purposes, rationalized by either a utility scheme or a complete preference ranking; that they deliberately consider a range of possible actions and their consequences; and that they choose an action based on its contribution to achieving these aims. This description requires that we focus attention on the agents' *goals* and *beliefs* and the *rules of choice* through which rational agents select one action from a range of alternatives.

We may begin with the problem of characterizing the goals of action, the goods that actions are designed to achieve. Individuals perform actions in order to acquire various things—income, leisure, education, and so on. And their actions impose costs on these agents: labor expended, wages forgone, risks run. To make rational decisions about various possible actions, then, it is necessary to have some way of weighing trade-offs between heterogeneous goods and bads because various goods and bads will commonly be produced by each possible choice. Is it worth it to me to give up an afternoon with my friends in order to hear an instructive philosophy lecture? If I have no way of comparing the goods associated with these two activities, then I have no basis for choosing between them.

Rational choice theorists use the concept of *utility* as a basis for comparing heterogeneous goods and bads or benefits and costs. A theory of utility is designed to provide a common measure for a variety of goods—income and leisure, nutrition and cost, intellectual challenge and social environment. The intuitive idea is that we can assign comparable values to heterogeneous goods because we do in fact manage to choose among them. The theory of utility is intended to formalize that capacity. The basic logical requirements for this theory are (1) that utility is a function that takes goods as a variable and specifies the value of the good to the agent as a result, (2) that a rational agent always prefers outcomes with greater utility, and (3) that the utility scale is continuous (so it is possible to add utilities).

We assume, then, that decisionmakers are able to assign utilities to all the goods that they value and that these utilities provide a basis for making choices among goods. For example, a prospective vacationer may judge that a trip to St. Tropez will produce better meals, worse beaches, and higher costs than a trip to Martinique. The decisionmaker needs a way of comparing the trade-offs of meals, beaches, and costs so that he or she can choose

the best vacation, all things considered. Utility theory offers a basis for doing just that (at a conceptual level, at least); it requires that the agent decide how much he or she would sacrifice in the quality of meals in order to improve the beach payoff, and so on for each of the goods in question. Notionally the agent might reason as follows: The meals at St. Tropez will produce a utility of 5 units compared to 3 units for Martinique; the beaches at St. Tropez produce 2 units, compared to 4 units for Martinique; and the cost of St. Tropez is -6 units, compared to a cost of -4 units for Martinique. This produces an overall utility of 1 unit for St. Tropez compared to 3 for Martinique—dictating the choice of Martinique over St. Tropez.

In some cases income is a suitable surrogate for utility, but it is not always so for it is reasonable to hold that income is subject to a law of diminishing marginal utility. That is, the benefit a worker derives from an increase in income from $10,000 to $15,000 is greater than the increase from $25,000 to $30,000. This implies that it may be rational for the worker to accept a more dangerous or unpleasant job to gain the first increase but not the second; the utility of the first $5,000 increment is greater than the disutility of the unpleasant job, whereas the disutility of the job is greater than the utility of the second $5,000 increment.

Several problems confront utility theory. How should we interpret the claim that "person p assigns utility u to outcome y"? Is this a psychological fact about the agent? Does it represent the amount of pleasure that the agent attaches to the outcome? Neither of these options has provided a plausible basis for the theory of utility. Instead, it is preferable to regard utilities as an abstract construct representing the value that the agent attributes to outcomes, permitting us to explain the choices and comparisons that the agent makes among them.

A second issue concerns the problem of "interpersonal comparisons" of utility. How are we to understand sentences like "p1 assigns the same utility to outcome y as p_2 does"? This is a particularly vexing problem if we assume that utilities are psychological magnitudes; it is less of a problem, however, if we regard utility as a theoretical construct in terms of which we can analyze agents' choices. Moreover, most applications of rational choice theory do not require interpersonal comparisons of utility because we are typically concerned with an actor's choices given his or her utility scale. (The problem of interpersonal comparisons arises in a serious way, however, in welfare economics, where the central task is to select policies that produce the greatest overall utility across a number of persons.)

An alternative approach to analyzing the agent's goals is to describe the agent's preference ranking of the outcomes rather than attempt to assign utilities to the outcomes. This approach is an *ordinal* framework (as compared to a cardinal utility framework). A preference ranking provides information concerning the agent's ranking of all pairs of outcomes, but it provides no information about *intensity* of preference. Let us understand the expression **"xPy"** to mean "the agent prefers x to y or the agent is indifferent between x and y." (Preference is thus conceived along the lines of the greater-than-

or-equal relation between numbers.) Suppose the range of options include (**a,b,c**) and the agent's preference rankings are: **aPc, cPb, aPb.** This is a *complete* preference ranking of the options in this sense: For each pair of alternatives **{x,y}**, it specifies whether **xPy** or **yPx.** And it is a transitive ranking in this se se: If **xPy** and **yPz,** then xPz. What a preference ranking does not offer, however, is information about how close together various choices are—information about intensity *of preference.* It might be, intuitively, that the agent's preference for **a** over **c** is very great, whereas the preference for c over b is slight, but a preference ranking cannot embody this information. Such information is, intuitively at least, relevant to decisionmaking. Fortunately it is possible to infer intensity of preference if we assume that agents have preferences between sets of outcomes specified probabilistically. Suppose that Jones prefers **a** to **b** and **b** to **c**. Now suppose that we offer him a series of choices between **b** and a lottery ticket with a fixed chance of winning a and the balance of winning **c**. There will be a probability **p** such that Jones is indifferent between **b** and the lottery ticket {**a** at probability **p; c** at probability 1–**p**} . Intuitively, this thought experiment can be understood as posing this question: How probable would a lottery ticket have to be in order to make it worthwhile to give up the certainty of **b** for the chance of gaining **a**? If Jones strongly prefers a to b and only slightly prefers **b** to c, then we would expect that the probability would be low. Let **k** be the probability at which the agent is indifferent between b and the lottery ticket. We can now assign notional utilities to **a**, **b**, and **c**: **U(a)=1, U(b)=k,** U(c)=0. The greater that **k** is, the closer together **a** and **b** are in Jones's preference space. This, then, is a technique for converting information about preference rankings into information about utilities. Therefore I will assume in what follows that it is possible to assign utilities to outcomes.

Probability

The theory of utility gives us a way of representing the goals of action. Now we need to consider the problems of risk and *uncertainty.* It is rarely possible to determine the outcome of an action with certainty; instead, in choosing a line of action, the agent must take into account the fact that there are multiple possible outcomes. The concept of risk refers to the common circumstance that a given action may have several possible outcomes with known probabilities, some of which are desirable and others undesirable. If I know that one out of ten plates of sushi are contaminated, then my choice of sushi for lunch is subject to risk: I have a 90 percent chance of enjoying my lunch and a 10 percent chance of food poisoning. Uncertainty refers to the fact that it may not be possible to determine the relative frequencies of outcomes. For example, if I know that some sushi is contaminated but do not know how common this problem is, then my decisionmaking is subject to uncertainty.

The central concept used in describing risk and uncertainty is that of the *probability* of an event or outcome. In general the probability of an event is an estimate of the likelihood of its occurrence, ranging between 0

and 1. An event with probability 0 is one that cannot occur; an event with probability 1 is one that is bound to occur. But what is the meaning of fractional probability values for a given event? There are two primary interpretations available: a frequency interpretation and a degree-of-belief interpretation. (The two interpretations are sometimes referred to as *objective* and *subjective* probabilities.) The frequency interpretation requires that we identify the universe of possible outcomes; the probability of a given outcome e is then the frequency of e within this universe of outcomes. For example, the probability of getting the ace of clubs in a bridge hand is .25—that is, one out of four randomly drawn bridge hands contains the ace of clubs. The other central interpretation construes a probability estimate as an indication of the strength of the agent's grounds for expecting the occurrence of the outcome, based on available evidence. When the weatherman judges that there is a 33 percent likelihood of rain, his statement rests upon the evidence available (the incoming low-pressure front), along with some rudimentary theory about the causal properties of the weather phenomena in question. (We may construe this as corresponding to the odds that the agent would accept in a wager concerning the event.) The subjective interpretation is most useful in discussion of uncertainty. In cases of uncertainty, we have no way of estimating relative frequencies of outcomes. We are therefore forced to assign equal a priori likelihood to each outcome—which is equivalent to saying that we have no greater reason to expect that **e** will occur than that any of the other possible outcomes will. (Discussion of these interpretations may be found in Glymour 1980.)

There is also a hybrid interpretation that relies on both these accounts. Here the judgment that "the probability of e is r" should be understood as representing *two* probabilities. (We may call this the *predicted frequency* interpretation of probability.) The probability claim itself can be understood as an estimate of the frequency of e within the universe of outcomes, and the degree of confidence that we have in the judgment is w (for warrant). Both r and w are values between 0 and 1, but there is no necessary relation between them. It may be that I have high warrant in believing that the incidence of failures in a nuclear power plant is low; in this case, w is high and r is low. For an example that runs in the opposite direction, suppose that the current theory of star formation implies that it is highly probable that the sun will burn out within one million years and that the evidence available for this theory is weak. The probability judgment that derives from this theory assigns a high probability r to the sun's burning out in one million years, but the warrant w that this judgment bears is low.

In general the frequency interpretation is preferable for scientific explanation; this is because we do not want to explain an event in the world on the basis of facts about our own states of mind (as is the case in the subjective interpretation). The difficulty is that for many events there is no straightforward way of computing the absolute incidence of the event in question. For example, suppose that it is held that there was a .33 probability of war between the United States and the Soviet Union at the time of the

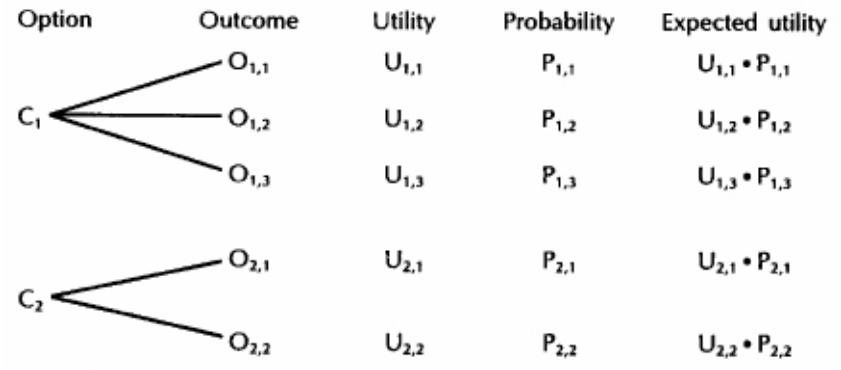

Fig. 3.1 Choices, outcomes, and probabilities

Cuban missile crisis. This is a nonrepeatable event; there is no existing universe of outcomes that can be used as a basis for frequency computations. Instead the frequency interpretation depends on counterfactual judgments: If the circumstances existing at the time of the crisis were rerun a large number of times, the incidence of war outcomes would be .33. Plainly this is an experiment that cannot be run, so our judgment that the probability of war was .33 must rest on other grounds—our theory of the causes of war. The predicted frequency alternative serves us best in this case: The meaning of the claim involves a hypothetical incidence of outcomes among possible alternatives, whereas *the warrant* for the claim depends on our theories of the causes of war applied to the particular circumstances of the missile crisis. Throughout, then, I will interpret probability judgments as estimates of relative frequencies, and I will set aside the problem of measuring the degree of warrant that these judgments possess.

Let us now consider a simple rational choice problem. The agent is faced with a range of alternative actions that may be performed, and each action has one or more possible outcomes with varying probabilities (the circumstances of risk and uncertainty). Figure 3.1 represents a simple example. The agent has two possible actions (C1 and C_2); C1 has three outcomes ($O_{1,1}$, $O_{1,2}$, and $O_{1,3}$), and C, has two outcomes **($O_{2,1}$** and $O_{2,2}$); and each outcome is associated with a payoff (Ui,j) and a probability (Pi,j). We assume, first, that the agent is able to assign values to the payoffs of each possible outcome. We may refer to these values as utilities. Second we assume that the agent can assign probabilities to each outcome; these probabilities may be construed as representing predicted frequencies of outcomes in repeated trials.

Decision *rules*

This analysis provides an abstract framework for analyzing the problem of rational decisionmaking. Now we must tackle the problem of articulating an appropriate decision rule. Return to the problem of choice described in

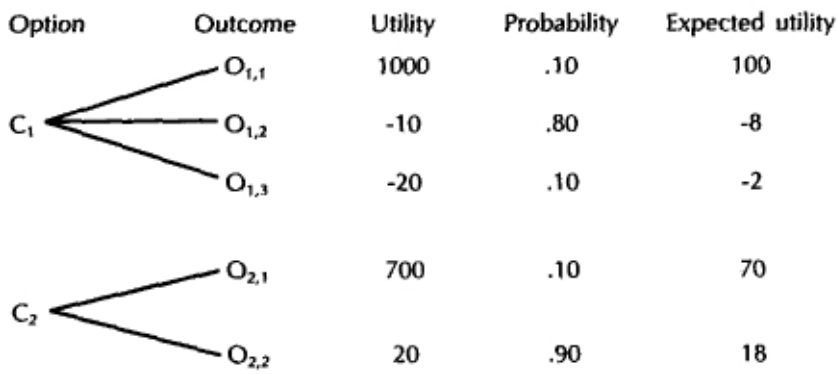

Fig. 3.2 A specific expected utility example

Figure 3.1, How should the agent decide what to do? One prominent basis for choice is the *expected utility* rule (sometimes referred to as Bayes' rule [Levi 1967:43-45]). In this approach the agent assigns a weighted value to each option that consists of the sum of the expected utilities for each of its outcomes (the utility of the outcome discounted by the probability of the outcome—$U_{i,j} * P_{i,j}$**).** The agent then chooses that outcome with the greatest expected utility. The advantage of this rule is that it leads to the greatest utility *when applied over a large number of choice situations.* If the problem of choice is one of deciding which lottery ticket to purchase and if the agent faces this choice frequently, the expected utility rule will lead to the highest possible winnings over time.

However suppose that the values of $U_{i,j}$ and $P_{i,j}$ are as described in Figure 3.2. In this example the expected utility of C_1 is 90, and that of C_2 is 88, so the expected utility rule would dictate the choice of C_1. However there is a 90 percent probability that the payoff for C_1 will be negative, whereas the payoff for C_2 is guaranteed to be positive (either 20 or 700). Finally suppose that this choice is a one-time opportunity, so that a loss today will not be evened out by future gains. Under these circumstances the expected utility rule does not seem to be a sensible rule of choice; it leads the agent to run a high risk of a loss when a gain can be guaranteed at only a small cost in the best case (by adopting C_2 over C_1.

Another rule that might be applied is called the *maximin* rule of choice. In this case the agent considers each alternative and identifies its worst outcome, then chooses that action that has the best worst outcome. (This rule leads the agent to maximize the minimum payoff received.) In the example of Figure 3.2, the worst outcome for C_1 is -20, and for C_2 it is 20; the maximin rule therefore dictates that the agent should choose C_2. The maximin rule is a "risk-aversive" rule; it protects the agent against catastrophic losses—even though it may also guarantee that the best achievable outcome will be lower than what might otherwise be gained.

These two rules differ in their treatment of risk and uncertainty, but each is a maximizing rule and requires that the decisionmaker choose the option

that optimizes with respect to a particular variable (expected utility or worst outcomes). Not all rational behavior depends on maximizing, however. Instead Herbert Simon has shown that much rational action derives from a decisionmaking process that he refers to as *satisficing* (Simon 1979). In this procedure the agent determines the minimal parameters that must be fulfilled in solving a problem. He or she then looks for a solution that satisfies these parameters and selects the first such solution. This process will not lead to the optimal solution to the problem, but it will produce a satisfactory one.

Satisficing behavior reflects an important constraint on rationality: the fact that there are information costs associated with the search for an optimal solution to a problem. If I want to eat the cereal that gives me the greatest nutritional payoff for the lowest possible cost, I must expend a good deal of effort evaluating all available cereals. There will be trade-offs between different nutritional parameters, so I will have to construct an appropriate metric assigning an overall nutritional value to each cereal. And finally I will have to balance cost and nutritional value. If, on the other hand, I want to eat a cereal that is "good enough," all I need to do is set a minimal standard of nutritional adequacy and a price standard and then choose the first cereal that I encounter that satisfies both requirements. (It might appear that satisficing choices maximize utility once we take information costs into account. However this is not quite accurate because to pursue a maximizing rule including information costs we would have to collect data on information costs and select an optimal solution in light of the new problem of choice. The satisficing approach dispenses with the need to collect additional information altogether once we have arrived at an acceptable solution.)

This approach to decisionmaking is particularly important in circumstances of complex choice-situations, in which there are many options and many possible outcomes. The cost of surveying all possible options and outcomes rapidly grows with an increase in the number of options; significantly, however, many real problems of choice do in fact involve large numbers of options. The satisficing rule thus appears to be an important basis for decisionmaking in complex real-life situations.

GAME THEORY AND THE PRISONERS' DILEMMA

Strategic rationality

The discussion to this point has analyzed rational choice on the assumption that the decisionmaker is confronted with a range of options with determinate outcomes (what Elster describes as "parametric" rationality [Elster 1983:74 ff.]). These cases involve the assumption that the outcomes are fixed by the properties of nature and that the decisionmaker's problem is simply to select one out of a menu of choices based on the probable consequences of each option. This framework covers a wide range of decision problems but not all. The most important class of cases that parametric rationality excludes are those in which outcomes depend on the deliberate choices of other

rational decisionmakers. This is the situation of *strategic* rationality, and problems of strategic rationality have a different structure than problems of parametric rationality. In particular the expected utility rule is no longer relevant as a rule of choice because outcomes are not probabilistic. In cases of strategic rationality, the payoff to the individual depends on the choices made by the other players. So each decisionmaker must consider the rational calculations of the others and choose that option that maximizes his or her payoff *given* the assumption that all the others make a rational decision as well.

Strategic rationality is particularly germane to social science because it bears on interactive social behavior: Individuals make choices based on their predictions about the actions other agents will perform, and the outcomes that individuals receive depend on the choices of other agents. This topic is the subject for investigation for several areas of rational choice theory, such as game theory and collective goods theory. In this section I will discuss the main ideas of game theory; in the next section I will turn to collective action theory.

Game theory is generally concerned with problems of strategic rationality—problems in which the rational decisionmaker must take into account the fact that the outcomes of various possible actions available to him or her are influenced by the choices made by other rational decisionmakers. Whereas the rational gambler chooses among alternative actions on the basis of the probabilities of win and loss that he or she assigns to each bet, the rational general must take into account both probabilities (e.g., concerning the weather) and the strategic rationality of his counterparts in the contending army. The opposing general is attempting to work out an optimal strategy given his understanding that the *opponent* is a *rational agent*; consequently each participant will act on the basis of assumptions about the other's intentions. This problem may look deeply intractable because A reasons that B reasons that A reasons that . . . , but the central finding of game theory is that there are optimal and stable solutions for several general classes of problems of choice of this sort.

Let us begin with the main ideas of two-person game theory. Game theory is premised on the assumption of rational self-interest and the theory of utility. Each "player" is assumed to have a set of private interests and a way of comparing the various possible outcomes in terms of their contribution to those interests. A *zero-sum* game is one in which each player's gain is exactly equal to the other player's loss; the sum of the two players' payoffs is zero. A non-zero-sum game is one in which the sum of payoffs for a given outcome may be positive (or negative, for that matter). An example of a zero-sum game is a bet on the toss of a coin; an example of a non-zero-sum game is an agreement between *a worker* and a capitalist to produce a good. It is evident that zero-sum games do not permit cooperation between the players because each player's gain is exactly offset by the other's loss. A zero-sum game is a game of pure competition. A positive-sum game, by contrast, *does* permit cooperation. For example, the winner may secure

		Column					
		$S_{2,1}$	$S_{2,2}$	.	.	.	$S_{2,m}$
	$S_{1,1}$	4,2	-1,3	.	.	.	5,-3
	$S_{1,2}$	6,3	2,0	.	.	.	-3,0
Row	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	$S_{1,n}$	-2.1	3,0	.	.	.	0,3

Fig. 3.3 Game matrix

the loser's cooperation by compensating him for his loss and still come out ahead. Thus a positive-sum game is a mixed game of competition and cooperation.

A strategy is a detailed rule of play for the whole of a game. It specifies the player's play for every possible move by the opponent at each stage of the game. (It is worth noting that this conception of a game strategy is extremely demanding; even in the game of checkers, the list of logically possible strategies for one player is impossibly long.) Each player is assumed to have a list of available strategies ($S_{i,j}$), each is assumed to know both his or her own list of strategies and that of the opponent (in a game of perfect information), and each is assumed to know the outcome for each player of a given pair of strategies.

There are two ways of representing a game. A game may be described in terms of its *game tree.* (This is termed as the *extensive form* of the game.) A game tree begins with the first player's options at the first move. For each of these options it specifies the options available to the second player and so on until the end of the game. Each complete branch of the game tree represents a pair of strategies for the two players. The advantage of the game tree is that it displays the game as a sequential series of plays by the two players. A finished game tree permits us to analyze the strategic situation of both players from the endstates backward. Each assumes that the other player is perfectly rational. At any stage of the game, the player can determine which set of options is available to the opponent on the next play. More generally earlier moves determine what sets of outcomes will be accessible later in the game. Because the opponent is assumed to be rational, the problem for each is to choose a strategy that forces the opponent to permit him to arrive at the best-worst payoff (an application of the maximin principle). There is a combinatorial explosion, however, that quickly threatens to overwhelm the analysis of any but the simplest of games; if each player has three choices at each play and if the game continues through five moves for each player, the total number of branches in the tree is 19,683 (3^9).

A game may also be summarized in the form of a "game matrix": a two-dimensional matrix listing player A's strategies in the rows and player **B's** strategies in the columns (Figure 3.3). (This is described as *normal form*

or strategic *form.)* Each entry in the matrix is an ordered pair that represents A's payoff and **B's** payoff for the selected pair of strategies. (The combinatorial explosion is equally significant in the case of normal-form descriptions of a game. For example, tic-tac-toe presents at least $9.7^{8}+5^{48}$ strategies to the first player.) A game *of perfect* information is one in which each player has full information about the strategies available to the other and about the state of the game at each stage of play. (That is, there are no hidden moves.) The central problem for two-person game theory, then, is to determine whether there are rational procedures for choosing strategies for games analyzed along these lines.

Let us begin with the analysis of two-person zero-sum games. The simplest strategic situation is the game in which each player has a dominant strategy—a strategy that is best for that player no matter what choice the opponent makes. In this case the player can effectively ignore the possible choices that the opponent may make; whatever my opponent does, my best strategy is fixed. In games in which each player has a dominant strategy, the outcome is easily determined: It is the intersection of the pair of dominant strategies. And if only one player has a dominant strategy, the problem of choice is also simple. If my opponent has a dominant strategy, then I know that he or she will play that strategy, and I should choose the strategy that gives me the greatest payoff on that assumption. In the more interesting cases, however, neither player has a dominant strategy; instead each must take into account the strategies available to the *opponent* and select a strategy accordingly.

Game theorists have shown that there are two classes of two-person zero-sum games. Some have a pure equilibrium point: a pair of strategies for players A and B with the property that—if these strategies are chosen—neither A nor B can improve the payoff by defecting to another strategy. (Such a position is also called a saddle point—an entry in a game matrix that is a maximum for one player and a minimum for the other.) That is, given that A chooses $S_{1,i}$), B can do no better than to choose $S_{2,j}$; given that **B** has chosen $S_{2,j}$, **A** can do no better than to choose $\mathbf{S_{1,1}}$**.** (This is sometimes referred to as a *Nash* equilibrium.) Under these circumstances both players have a best available strategy, and the game is solved. How can we determine whether a given game has an equilibrium point? Here the maximin rule described above is the appropriate tool of analysis for each player. (I will refer to the players as "Row" and "Column.") Row should rank his strategies according to their worst outcomes and provisionally choose that strategy $S_{1,i}$ with the best worst outcome. Now he should consider what options are available to his opponent: If Column knew that Row is playing $S_{1,j}$, what strategy would he choose? On the assumption that Column would choose $S_{2,i}$, could Row improve his payoff? If he could, then $S_{1,i}$ does not provide an equilibrium point; if he could not, then $\{S_{1,j}, S_{2,i}\}$ is an equilibrium point. Games in which such an equilibrium exists likewise have an optimal solution: Each player should choose a strategy that falls on an equilibrium point. If there is a saddle point, then each player can do no better than choose a strategy that leads to this saddle point.

The second class of games is more difficult. These are games without a saddle point; consequently no single pair of strategies represents an equilibrium. (If Row chooses S1,i on the assumption that Column will play $S_{2,j}$, then Column can improve his payoff by choosing another strategy. Row, foreseeing this possibility, determines not to play Sm.) Game theorists have shown that these games too have a solution for both players. In this case, however, the solution is *a mixed* strategy: a distribution among several different strategies determined by a fixed set of probabilities assigned to them (to be applied using a randomizing process). The advantage of a mixed strategy is that it makes it impossible for my opponent to exploit knowledge about what I will do. If he knows that I must play $S_{1,j}$, he can choose the best response available on that assumption. But if he knows that I will choose randomly among $\mathbf{S}_{1,j}$, $\mathbf{S}_{1,k}$, and $S_{1,l}$, then he must be prepared for each of these strategies.

The analysis up to this point is restricted to zero-sum games. However many cases of strategic interaction are not zero-sum; instead many games produce outcomes in which both parties may be better off if they cooperate. A game *of pure cooperation is* the polar case. In this situation the optimal outcome for both players is possible if they properly coordinate their strategies. There is no conflict of interest between the parties; each is concerned only to coordinate with the other. (An example of this is the problem of locating a friend in a crowded stadium. It does not matter whether the friends meet at the ticket booth or the 50-yard line, as long as they both arrive at the same place.) The more interesting case is that in which there is both harmony of interest and conflict of interest between the players. In such a case both do better by coordinating with each other, but each prefers some of the cooperative outcomes to others. Thus there is a conflict of interest between the players over which of the cooperative outcomes will be selected.

This case is of particular interest in the social sciences. Given that the game is non-zero-sum, a negotiated solution is possible. (In a zero-sum game there is no overlap of interest between the two players that would permit a negotiated solution.) Here certain outcomes are preferred by all players over other outcomes, and if players are permitted to communicate with each other, they may be able to reach an agreement that enables them to coordinate their choices and arrive at one such outcome. Game theorists have tried to analyze the conditions that affect what the bargaining solution will be, based on the payoffs to the parties. Intuitively the general conclusion is that, if I am the party with the most to lose, I will be forced to accept a bargained solution that favors my opponent for he can use his "threat advantage" to reason that failure to reach agreement will hurt me more than him. (See Shubik 1982 for an extensive discussion of the large literature on bargaining theory.)

The prisoners' dilemma

These are the basic notions of game theory. And—as game theorists themselves point out explicitly—the theory has few direct practical appli-

	Cooperate	Defect
Cooperate	1,1	-2,2
Defect	2,-2	-1,-1

Fig. 3.4 Prisoners' dilemma

cations because of the impossibly strenuous assumptions it makes about each player's knowledge and computational abilities. Along the way, however, the game theorists have analyzed several simple games that have surprising properties. Central among these is the prisoners' *dilemma.* This is a nonzero-sum game that models a number of common strategic situations. Consider the game matrix in Figure 3.4. Each player is faced with two possible strategies: cooperate and defect. The payoffs to each are as indicated; A's strategies are listed on the left side, and **B's** are listed across the top; A's payoff is the first quantity, and **B's** payoff is the second quantity. If both choose to cooperate, then both receive 1 unit; if both defect, both lose 1 unit. Finally, if one cooperates and the other defects, the defector gains
2 units and the cooperator loses 2 units.

If we now analyze this game according to the assumptions of rational self-interest outlined above, we will see immediately that it has an equilibrium point because each player has a dominant strategy. The dominant strategy is defection: Each sees that he is better off defecting regardless of whether the opponent defects or cooperates. And the equilibrium point is the pair of defecting strategies—with a loss of 1 unit for both A and **B.** Both players prefer the cooperate-cooperate outcome to the defect-defect outcome, but they are unable to arrive at this outcome through rational decisionmaking. Here we have arrived at something like a paradox of collective rationality. Each player chooses rationally, each selects a strategy that maximizes his own outcome, and the net result is an outcome that is worse for both than another possible outcome (joint cooperation). It would seem, then, that individual rationality in this case leads to collective harm.

There are many instances of social behavior that appear to *embody* the structure of the prisoners' dilemma—for example, arms races, the breakdown of price-fixing agreements, and the failure of cooperative practices. In each case participants have a collective interest in a cooperative solution that is undermined by the cost-free incentive to defect. Prisoners' dilemmas thus involve the role of *trust:* If parties to a cooperative agreement trust that other participants will keep the agreement and if each participant has a normative motivation to keep fair agreements, then prisoners' dilemma situations can be overcome.

The situation of a prisoners' dilemma changes if the situation of choice is a repetitive one.[3] Here the chief finding may be summarized rather simply: Defection is no longer the optimal strategy for each player when each knows that he confronts an open-ended series of prisoners' dilemma decisions with a given opponent.[4].' Each player can foresee that defection on the first play—even if it gains a one-time advantage over the opponent—will lead the

opponent to defect on the second play and will result in a stable run of "defect" plays by each player for the whole series of games. Each can see that both players lose on this scenario; consequently it is rational to make a tacit "offer" to cooperate through playing "cooperate" on one play and seeing if the opponent reciprocates on the next. Robert Axelrod (1984) and Michael Taylor (1987) have analyzed the structure of cooperation from the point of view of prisoners' dilemmas; see Example 3.4 below for Axelrod's analysis. These arguments show that conditional cooperation is a rational strategy in repeated prisoners' dilemma situations.

Applicability of game theory to empirical social science

It is reasonable to ask, in at least a preliminary way, to what extent game theory is relevant to empirical social science. The technical apparatus of game theory is probably less useful than the basic ideas that game theory provides: strategic rationality, the prisoners' dilemma, reasoning about the choices of others in a circumstance of interactive outcomes, and bargaining and coalitions. The technical achievements of game theory—that various classes of games are in principle solvable—are of questionable relevance. Suppose that a peasant community and its lord in a particular historical context are in a conflict with a game structure that locates it within a class of games T. Suppose further that some axiomatization of game theory shows that T is solvable using a particular mixed strategy. Nothing follows from these facts for the behavior of the participants, even if we assume that they are rational, for the participants do not know that T is solvable, and, in any case, they lack the mathematical machinery for solving T. The fact that they are rational does not imply that they will act in accordance with the requirements of a fully developed scheme of strategic rationality. In fact, if their actions do conform to the solution to T, we have an even harder problem—explaining how this fortuitous outcome emerged. This situation parallels arguments for the applicability of game theory to evolutionary biology in the form of the concept of evolutionarily stable strategies. It is held that species may evolve genetically determined "mixed strategies" in which individuals are programmed to alternate strategies in a fixed ratio and that the ratio is that required by the solution to the game in which the species finds itself in a given environment. (See Elster 1982 and Dawkins 1976 on these applications.) The biological explanation would go along these lines: Those subpopulations that accidentally hit the right strategic mix have an advantage over those that do not.

How would this affect the peasants/lord game discussed above? Not at all. The behavior of peasants and lords is not genetically programmed but rather intentional and rational. If they do not possess the machinery of game theory, they could only hit the optimal mix of strategies through trial and error, not through rational calculation.[5]

The general framework of analysis provided by game theory, however, is useful for social science explanation. Consider Example 3.4. Here Robert Axelrod uses some nontechnical elements of two-person game theory to

Example 3.4 Cooperation and repeated prisoners' dilemmas

In World War 1 the violence of trench warfare was often reduced by apparent unofficial truces by units on both sides. Each side would continue to fire its weapons but without inflicting much damage on the other. Robert Axelrod explains this "live and let live" strategy in terms of the phenomenon of *reciprocity* (strict conditional cooperation). He argues that rationally self-interested agents will find it in their self-interest to cooperate conditionally with other agents in circumstances where each side has something to gain from cooperation and something to gain from defection: The short-term gains of defection are more than offset by the long-term gains of cooperation. Axelrod's model of cooperation derives from study of repeated prisoners' dilemmas. He shows that the structure of the prisoners' dilemma changes in an open-ended series of plays of the game. Conditional cooperation ("tit for tat") is the best strategy for each player and the most robust over a wide variety of contexts. "Tit for tat" opens with cooperation and then plays whatever its opponent played on the previous move—that is, it responds cooperatively to cooperation and immediately punishes defection with defection. Axelrod identifies a set of conditions under which cooperation (strict reciprocity) is the optimal strategy for each player. Players must first be able to recognize and reidentify their opponents from one play to the next, and they must be able to remember the opponents' previous history of play. These conditions are necessary to make the cooperator selectively responsive to different strategies. Then players must judge that the probability of future interaction with the opponent is sufficiently great to justify weighing future gains from cooperation against present gains from defection. Under these circumstances Axelrod shows that the optimal strategy for each individual when confronted with opportunities for cooperation with others is conditional cooperation. Axelrod holds that the "live and let live" process found in trench warfare is explained as rational behavior making use of the strategy of conditional cooperation on both sides.

Data: examples of cooperative behavior, game theoretic analysis of repeated prisoners' dilemmas, historical data from World War I

Explanatory model: explanation of patterns of cooperative behavior as the result of the rational self-interest of each of the players

Source: Robert Axelrod, The Evolution of Cooperation (1984)

account for cooperation among rational agents. This analysis is useful in explaining a range of social phenomena, from tipping behavior to the practice of aiming high in trench warfare. In this instance we have a situation that embodies a repeated prisoners' dilemma between the two sides. On any given occasion each side prefers the outcomes in this order: unilateral shooting, joint nonshooting, joint shooting, and unilateral nonshooting. (That is, each side would prefer to impose harm on the enemy without cost to itself.) If each unit encountered **an** enemy only once, we would expect that each side would shoot. Given the situation of trench warfare, however, in which opposing units face each other over an open-ended series of opportunities for conflict, the strategy of conditional cooperation is superior

to noncooperation for each party; each is better off if it continues to cooperate in response to previous cooperation by the enemy. (We should also expect this pattern of cooperation to break down as one side or the other comes closer to withdrawal from the front.)

COLLECTIVE ACTION THEORY

Let us turn now to another important area of applied rational choice theory: the theory of collective action. A generation of economists writing on public goods problems have shown that there is conflict between private rationality and collective action: A group of rationally self-interested individuals will not act effectively in pursuit of public goods (goods that are indivisible and nonexcludable—for example, clean air and water). In a classic work Mancur Olson (1965) advanced a theory of group behavior that drew certain counterintuitive conclusions. There is a long-standing tradition of thought that took it as self-evident that groups and organizations would act collectively in pursuit of the common interest of the group. Olson showed, however, that this assumption commits something akin to a logical fallacy because groups consist of individuals who make independent decisions. Consequently it is not sufficient to show that an action would serve the group's interest if all or most members of the group were to perform it; it is necessary to show in addition that all (or most) individuals in the group have a rational interest in acting in that way. (Russell Hardin uses the term "fallacy of composition" to describe the error [Hardin 1982:2].) In fact Olson argues that in the most common circumstances a group will not act effectively in pursuit of common interests. Rather, if we assume that a group is composed of rational agents concerned with maximizing private interests, Olson shows that each member will have a rational incentive to take a "free ride." Each potential contributor to the public good will choose to become a free-rider and hope that other members of the group will make the contrary decision.

Assume that a group is composed of rational individuals who have a common interest—an outcome that would benefit each of them if it were to occur. Individuals are motivated by self-interest, described by a consistent set of utilities. Every individual has a range of private interests and chooses among available actions according to the costs and benefits that each presents in terms of those private interests. Assume that a common interest is a good whose attainment would improve every individual's welfare, according to his or her own private scheme of interests. (That is, there is no conflict of interest over the attainment of the good; every member of the group would prefer the presence of the good to the absence of the good.) Assume that this common good is a public good—a good that, if it is available to any member of the group, "cannot feasibly be withheld from the others in the group" (Olson 1965:14). (That is, a public good is characterized by nonexcludability.) Finally the collective action of the group is the action that is expected of each member in order to achieve the common good. The problem of collective action is this: Under what circumstances will a group succeed in acting in concert to bring about its common interest?

Consider an example that embodies these assumptions. Let the group be an association of mail-order merchants, let the common good be a decrease in the postage rate on the mailing of catalogs, and let the collective action in question be a contribution to a lobbying fund designed to get Congress to write appropriate legislation. Assume, further, that the lobbying effort is almost certain to be successful if funded at a sufficiently high level—say, 90 percent compliance with each member donating $1,000. Finally, assume that the savings in postage that each member would realize would average $800 per year, over a predicted time frame of five years. The good in question in this example is a common good; each member would benefit from the decrease in postage rates. Further, it is a public good; it is not possible to exclude noncontributors from the benefits of lower postage rates. Finally, the individual rationality assumption is satisfied if we simply assume that merchants decide whether to contribute strictly according to their individual costs and benefits of contribution or noncontribution.

We are now ready to consider what I will refer to as Olson's "theorem of collective action." "In a large group in which no single individual's contribution makes a perceptible difference to the group as a whole, or the burden or benefit of any single member of the group, it is certain that a collective good will not be provided unless there is coercion or some outside inducements that will lead the members of the group to act in their common interest" (Olson 1965:44). His argument reduces in large part to the following point: "Though all of the members of the group . . . have a common interest in obtaining this collective benefit, they have no common interest in paying the cost of providing that collective good. Each would prefer that the others pay the entire cost, and ordinarily would get any benefit provided whether he had borne part of the cost or not" (Olson 1965:21). The theorem follows from two points: the fact that rational individuals make their decisions based on private interests and the fact that the common good is nonexcludable. Given nonexcludability, individuals can reason that the good will either be achieved or not achieved independent from their own choices of action. With either outcome, personal interests are best served by not contributing. If the good is achieved, they will enjoy the benefits without the cost of contribution. If it is not achieved, then the individuals are spared the cost of contribution. Each member will thus decide not to contribute, and the good will not be achieved—thus the "theorem of collective action."

This problem of collective action is referred to as the "free-rider" problem: Rationally self-interested individuals are under an unavoidable incentive to take a "free ride" in circumstances of collective action—that is, to refrain from contribution and hope that others make a contrary choice.

We might informally test this result against the assumptions of our example above. Assume a representative merchant has just received the request for a contribution to the lobbying fund, reminding him that the association determined that this course will best serve the common interest. He has two choices: to contribute or not to contribute. And there are two possible outcomes: successful collective action and unsuccessful collective

	Success	Failure
Contribute:	3000	-1000
Don't contribute:	4000	0

Fig. 3.5 Collective action payoffs

action. These choices are represented in Figure 3.5. This table of outcomes shows that the merchant has a best strategy available regardless of the success or failure of the joint enterprise: noncontribution. This strategy leads to $4,000 versus $3,000 in the event of collective success; it leads to $0 versus -$1,000 in the event of collective failure. Therefore our representative merchant elects to refrain from contribution. But each participant is faced with the same scheme of costs and benefits. Therefore none contributes, the lobbying effort fails, and the good is not achieved.

Olson qualifies his analysis in two ways. First he distinguishes between large and small groups and shows that small groups are sometimes "privileged"; individuals in such small groups may derive enough benefits from the supply of the public good that it is individually rational to purchase the good. Large groups, however, are "latent": They normally do not succeed in undertaking collective action. Second he points out that groups may be able to arrange a schedule of in-process benefits or penalties that are sufficient to change the individual's rational calculus.

Russell Hardin shows that Olson's analysis of group size is too simple, however, and that a more complete analysis proves that size is relevant in other respects as well. In particular Hardin shows that more relevant than absolute size is the ratio of benefits to costs and the extent of stratification of benefits within the group (Hardin 1982:40 ff.). If the benefit-to-cost ratio is sufficiently high, there may be a subgroup within the larger group that would benefit from the collective good even if it provided the whole funding of the collective project. "Let us use k to designate the size of any sub-group that just barely stands to benefit from providing the good, even without cooperation from other members of the whole group" (Hardin 1982:41). Hardin shows that it is the size of k rather than the absolute size of the group that influences the feasibility of collective action. Suppose that a thousand people would benefit from extending road service to a remote village and that benefits are unequally distributed. Most people would save $100 a year on the cost of hiring a donkey to convey them to the city, but a small group of ten merchants would gain $5,000 a year in increased trade. Finally suppose the cost of the road is $10,000. The benefits to the 990 ordinary villagers are $99,000-much greater than the cost of the road. But, for reasons deriving from Olson's analysis, it will be difficult to secure cooperation from this group. The benefits to the ten merchants are $50,000, so it is in their interest to fund the whole cost of the road rather than have the project fail. Moreover this is a small enough group that we may expect that it will succeed in implementing this collective effort.

This case no doubt strikes the reader as closely related to the prisoners' dilemma sketched above. And in fact Russell Hardin argues that the problem

of collective action is formally equivalent to the n-person prisoners' dilemma (Hardin 1982:25-28). Both the prisoners' dilemma and the collective action theorem have apparently paradoxical consequences for group rationality. (They represent "the back of the invisible hand," in Hardin's felicitous phrase.) Both results appear to show that groups composed of rational individuals will be incapable of acting to secure collective benefits—even when all participants can rehearse the full story of Olson's argument and the prisoners' dilemma. And the only solutions that seem to be available for these problems in their most abstract form are either irrational conduct (choosing a less-than-optimal strategy) or coordination under coercive conditions (in which individuals can commit themselves not to defect from the collective action).

The theory of collective action provides the basis for the explanation of a wide variety of social behavior: strikes, the success or failure of rebellion, and the instability of price-fixing agreements. In Example 3.5, Allen Buchanan uses the collective action problem to explain worker passivity in the face of opportunities for revolutionary action.

CRITICISMS OF NARROW ECONOMIC RATIONALITY

This completes my treatment of the main tools of rational choice theory. In this final section I return to the issue with which we began: the specification of the notion of individual rationality. A number of writers have offered criticisms of the conception of individual rationality at work here, on the ground that it is insensitive to features of human action and deliberation that are in fact quite central.

Some authors have criticized various aspects of the theory of narrow economic rationality. Particularly important among these is A. K. Sen's critique. Sen—himself an economist of the first rank—criticizes the assumption of pure self-interest that is contained in the standard conception. "The purely economic man is indeed close to being a social moron" (Sen 1982:99). Against the assumption of self-interested maximizing decision-making, Sen argues for a proposal for a more structured concept of practical reason, one that permits the decisionmaker to take account of *commitments.* This concept covers a variety of nonwelfare features of reasoning, but moral principle (fairness and reciprocity) and altruistic concern for the welfare of others are central among these. Sen believes that the role of commitment is centrally important in the analysis of individuals' behavior with regard to public goods. For example, he suggests that the voters' paradox may be explained by assuming that "voters are not trying to maximize expected utility, but ... to record one's true preference" (Sen 1982:97). And he draws connections between the role of commitment and work motivation. "To run an organization entirely on incentives to personal gain is pretty much a hopeless task" (Sen 1982:98). He argues, therefore, that to understand different areas of rational behavior it is necessary to consider both utility-maximizing decisionmaking and rational conduct influenced by commitment; furthermore,

Example 3.5 Revolutionary motivation

Marxist theory predicts that workers have objective class interests that make it rational for them to support revolutionary movements to overthrow capitalism. Marx writes, "The proletarians have nothing to lose but their chains. They have a world to win" (Marx and Engels 1848/1974:98)**.** But proletarian activism and revolution are the exception, not the rule, among working-class groups throughout the world, Why is this so? Allen Buchanan explains this phenomenon by accepting the point that workers have a collective interest in revolution but pointing out that revolution is a "public good" for members of the working class. Buchanan argues, "Even if revolution is in the best interest of the proletariat, and even if every member of the proletariat realizes that this is so, so far as its members act rationally, this class will not achieve concerted revolutionary action" (Buchanan 1979:63). Any worker will be able to enjoy the benefits of socialism whether he has contributed to the revolution or not. Therefore rational workers elect to become free-riders. As a result working-class collective action is infrequent. Thus Buchanan derives working-class passivity from three assumptions: (1) workers have a group interest in revolution, (2) workers are individually rational decisionmakers, and (3) rational decisionmakers are usually ineffective at securing collective action. Therefore the working class is generally incapable of mounting collective action in support of its interests.

Data: historical patterns of working-class political behavior
Explanatory model: application of the theory of collective action to a hypothetical group of rational proletarians
Source: Allen Buchanan, "Revolutionary Motivation and Rationality" (1979)

---

it is an empirical question whether one factor or the other is predominant in a particular range of behavior (Sen 1982:104).

Sen's arguments show that there are good analytic and empirical reasons for judging that much actual human behavior is not explicable on the basis of a simple utility-maximizing scheme. This finding might lead us to suppose that human beings are typically not rational or it might lead us to question the concept of rationality associated with the standard conception. Sen suggests the latter course and proposes that we attempt to build a more structured concept of practical reason that permits us to take account of moral, political, and personal commitments as well as concern for welfare. Moreover he shows that the former cannot be subsumed under the simple concepts of utility-maximizing or preference rank-ordering. (Sen's main contributions are contained in "Rational Fools" and "The Impossibility of a Paretian Liberal" in Sen 1982.) Thus he holds that an adequate theory of rationality requires more structure than a simple utility-maximizing model would allow; in particular it must take account of moral principles and commitment.

This argument suggests that the concept of rationality must incorporate normative principles in some way. How might this be done? Recent work in moral philosophy offers some insight into this problem. Various moral

philosophers have argued that practical rationality is more comprehensive than narrow economic rationality. Thus Thomas Nagel provides a series of arguments to the effect that rationality requires altruism—recognition of the reality of the interests of others and a direct willingness to act out of regard for those interests (T. Nagel 1970). The egoism assumption is neither mandatory nor plausible as a basis for rational choice analysis. Instead it is perfectly consistent to postulate that individuals define a range of goals, from narrow self-interest to the interests of the family to the interests of more encompassing groups, and choose their actions according to the degree to which various alternatives serve this ensemble of interests. All that the rational choice requires is that these be *individual* goals—that is, goals established and pursued by individual agents. But the content of the goals may be other-regarding. It is the structure of means-end rationality rather than the particular character of the ends that individuals pursue that is essential for the rational choice approach.

This line of thought directly addresses the egoism assumption of the standard conception. It does not, however, do quite enough for it does not give us a way of incorporating the idea of moral principles (or other normative requirements) into the decisionmaking process. But other recent moral philosophers have outlined the sort of structured decisionmaking process necessary to take account of the role of principle in decisionmaking: The decisionmaker can combine a set of side constraints on action (normative commitments, in Sen's terms) as well as a set of goals (personal interest, social goals, the welfare of others, etc.).

In particular a number of philosophers have attempted to incorporate the idea of *fairness* into the concept of rational decisionmaking.[6] A reason for my performing an act is that I benefit from widespread performance of this sort of act, and I recognize that fairness requires that I pay my share of the cost of these public benefits. John Rawls's *A Theory of Justice* represents an extended argument to the effect that there are principles of justice that should regulate the just society, derived from the principle of fairness. His construction is at some distance from our primary concern because he is concerned with global features of justice and we are concerned with individual rationality. But the kernel of Rawls's construction is relevant here: If individual rationality involves evaluating alternative lines of action in terms not only of the costs and benefits of each alternative but of the fairness of each alternative, then we have arrived at a structured concept of rationality. And it is a concept that involves the imposition of side constraints on the decisionmaking process. A more structured decisionmaking process is necessary to take account of the role of principle in decisionmaking: The decisionmaker can combine a set of side constraints on action as well as a set of goals. And the decisions he or she arrives at will be a complex function of constraints and goal-maximizing actions.

How do these findings relate to our central concerns? First they suggest that the narrow conception of economic rationality is not a comprehensive theory of practical reason because it fails to consider certain features of the

decisionmaking process that are intuitively crucial in some contexts. Further these considerations suggest an alternative model of the decisionmaking process that promises to be a more adequate analysis of the concept of human rationality. Moreover this richer conception of practical reason promises to offer a new set of solutions to different classes of collective action problems: If individuals are altruistic to some degree (that is, responsive to the interests of others) and if they are principled (that is, moved by considerations of fairness, reciprocity, or justice), then they will be practically motivated to act differently, when confronted with occasions for collective action, than the theory of collective action predicts.

These findings have direct import for the applicability of rational choice models in social explanations. For example, consider the problem of free-riding and public goods problems. Once we consider a more complex theory of practical deliberation, formal arguments predicting the emergence of public goods problems in real social groups will be found to be misleading. On a more complex and more empirically adequate account of practical reason, altruism, cooperation, and reciprocity are rational choices; therefore we would expect a social group consisting of rational individuals to show marks of cooperation and altruism.

We must be careful not to draw an overly strong conclusion, however, for no one would maintain that human beings are indifferent to private welfare. Indeed generally speaking it would seem reasonable to assume that each decisionmaker places a high priority on personal and familial welfare; human beings generally do not behave like impartial utilitarians. This finding suggests that human behavior is the result of several different forms of motive, such as self-interest and altruism, and several different types of decisionmaking processes, including maximizing and side-constraint testing. (See Margolis 1982 for an attempt to formalize some of these contrasts.) And to the degree that self-interest and maximizing behavior are prominent in a particular type of circumstance, the collective action theorem will be empirically significant. These criticisms, then, do not discredit the rational choice approach; rather they suggest the need for further development of the theory of individual rationality.

CONCLUSION

This chapter has surveyed the foundations of the rational choice approach to social explanation. The general idea is to explain specific social phenomena as the aggregate result of large numbers of rational persons making choices within a specific social and natural environment. What gives social content to this approach is the level of detail provided about specifics of the social environment. So, for example, rational persons within a traditional peasant society may show substantially different patterns of behavior from those of persons in modern industrialized societies. And these differences may derive not from differences in the psychology or agency of the persons involved but from substantive differences in each group's environment of choice.

The rational choice approach underlies several important research programs in social science that will be considered in greater detail below. The *public choice* paradigm attempts to explain social and economic behavior of persons in non-Western societies on the basis of fairly narrow assumptions about individual rationality. This paradigm forms the basis of work in economic *anthropology.* And the rational choice paradigm has close affinities with materialist *explanation* and Marxist theory, for materialists and Marxist social scientists attempt to explain aggregate social structures as the result of rational individuals pursuing their material interests.

Before we turn to these applications of the rational choice approach, however, we must consider a powerful line of criticism against this approach—the view that social science requires *interpretation* of culturally specific norms, values, and meanings. This view suggests that the rational choice framework is fundamentally flawed because it attempts to abstract from the culturally specific content of agency and replace it with an abstract, universal model of rationality. In the next chapter, then, we will consider the interpretive paradigm of social explanation.

NOTES

1. For a brief but clear discussion of this type of theory of rationality, see Philip Pettit, "Rational Man Theory," in Hookway and Pettit, eds. (1978). Von Wright (1971) provides a more extensive analysis of rational-intentional explanations. My *Under*-standing Peasant China (1989) explores the application of this model to China studies.

2. For a useful discussion of "thin" and "thick" theories of rationality in area studies, see Michael Taylor's useful essay, "Rationality and Revolutionary Collective Action," in Michael Taylor, ed. (1988). This collection provides a number of strong examples of the rational choice approach in application to area studies.

3. Particularly important are Axelrod (1984), Rapoport and Chammah (1965), Hardin (1982), and M. Taylor (1976).

4. The qualification of open-endedness is important. If the series ends at the hundredth game, then each party foresees that the other will defect on the last game. But if the opponent is determined to defect on the hundredth game, then the player should defect on the ninety-ninth game and so forth back to the first game. See Hardin (1982:146 ff.) on this paradox.

5. Kenneth Oye provides a thoughtful consideration of the relevance of game theory to applied social science in his introduction to Cooperation *Under* Anarchy (1986). He writes, "The equilibrium solutions identified by formal game theorists may stabilize convergent expectations among mathematicians, but unless equilibria can also be reached through 'alternative less sophisticated routes,' such solutions may have little influence on international outcomes" (Oye, ed. 1986:2).

6. See, for example, the extensive literature on utilitarianism and fairness (Regan 1980, Griffin 1985, and Harsanyi 1985).

SUGGESTIONS FOR FURTHER READING

Axelrod, Robert. 1984. *The Evolution of Cooperation.*
Becker, Gary. 1976. *The Economic Approach to Human Behavior.*
Bonner, John. 1986. *Introduction to the Theory of Social Choice.*

Elster, Jon. 1979. *Ulysses and the Sirens.*
Elster, Jon, ed. 1986. *Rational Choice.*
Rapoport, Anatol. 1966. *Two-Person Game Theory.*
Schelling, Thomas C. 1978. *Micromotives and Macrobehavior.*
Sen, Amartya. 1987. *On Ethics and Economics.*
Shubik, Martin. 1982. *Game Theory in the Social Sciences: Concepts and Solutions*
Simon, Herbert A. 1983. *Reason in Human Affairs.*
Taylor, Michael. 1987. *The Possibility of Cooperation.*